(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2005 年10 月13 日 (13.10.2005)

PCT

(10) 国際公開番号
WO 2005/096208 A1

(51) 国際特許分類[7]: G06F 19/00, 17/30

(21) 国際出願番号: PCT/JP2005/006397

(22) 国際出願日: 2005 年3 月31 日 (31.03.2005)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2004-108456 2004 年3 月31 日 (31.03.2004) JP

(71) 出願人 (米国を除く全ての指定国について): 株式会社バイオシンクタンク (BIO-THINK TANK CO., LTD.) [JP/JP]; 〒1130033 東京都文京区本郷三丁目３２番６－７０３号 Tokyo (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 森下真一 (MORISHITA, Shinichi) [JP/JP]; 〒1790073 東京都練馬区田柄１丁目１９－７ Tokyo (JP). 山田智之 (YAMADA, Tomoyuki) [JP/JP]; 〒1130033 東京都文京区本郷５丁目１４－８ テラス本郷３０２号室 Tokyo (JP).

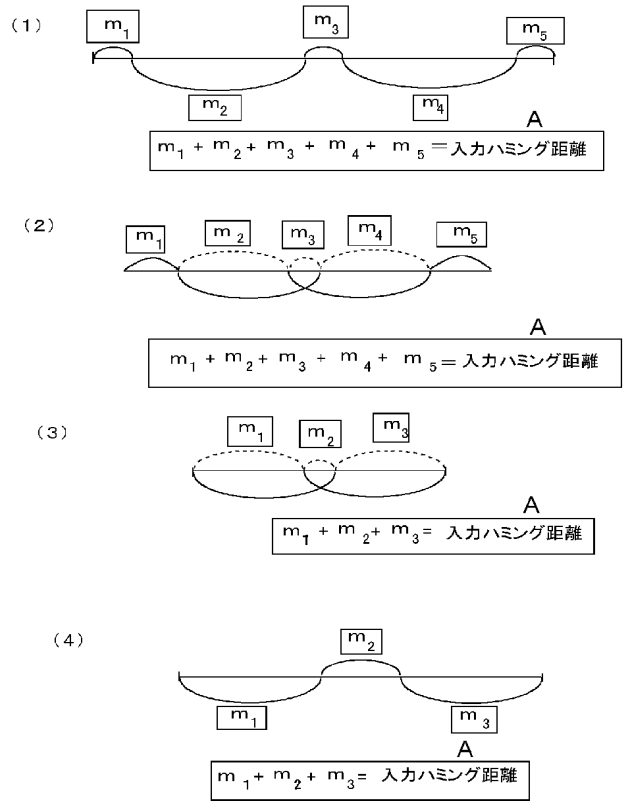
(74) 代理人: 工藤一郎 (KUDO, Ichiro); 〒1000006 東京都千代田区有楽町１－７－１有楽町電気ビル（南館）９階 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE,

[続葉有]

(54) Title: BASE SEQUENCE RETRIEVAL APPARATUS

(54) 発明の名称: 塩基配列検索装置及び塩基配列検索方法



A...INPUTTED HAMMING DISTANCE

(57) Abstract: An apparatus, method, etc. that in designing of the base sequence of, for example, siRNA, realize high-speed retrieval any genes containing analogous base sequences without omission. Accordingly, retrieval is carried out in such a manner that two partial sequences of given length and any extra part are identified from inputted base sequences, and that hamming distance being the number of corresponding bases incompatible with each other is divided and assigned to the partial sequences and extra part and out of the two partial sequences, one with an assigned number not greater is selected and retrieved.

(57) 要約: ｓｉＲＮＡなどの塩基配列を設計する場合に、類似する塩基配列を含む遺伝子を漏れなく高速に検索する装置及び方法などを提供する。このために、入力された塩基配列から所定の長さの二つの部分配列とその余の部分とを特定して、対応する塩基が適合しない数であるハミング距離を、それらの部分配列とその余の部分とに分割して割り当てて、２つの部分配列のうち、割り当てられた数が大きくないほうを選択して、検索を行なうようにする。

WO 2005/096208 A1

SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書

２文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明 細 書

塩基配列検索装置及び塩基配列検索方法

技術分野

[0001] 本発明は、遺伝子情報を表わす遺伝子塩基配列を検索する装置及び方法に関する。

背景技術

[0002] ワトソンとクリックとによるDNA(Doexyribo Nucleic Acid)の構造の解明に基づき、塩基配列に基づく遺伝子情報の研究が発展している。DNAは、アデニン(A)、シトシン(C)、グアニン(G)、チミン(T)の塩基のいずれかを含むヌクレオチドが並んでいる構造を持ち、細胞の核の中では、通常、AとT、GとC、の結合により、二重らせんの構造となっている。遺伝子を表現するDNAのヌクレオチドの配列(以下、「遺伝子塩基配列」と呼ぶ)が、RNA(Ribonucleic Acid)に転写され、スプライシングを経て、mRNA(messenger RNA)が生成され、たんぱく質の合成がされることが知られている。RNAは、D－リボースを糖成分として、アデニン(A)、シトシン(C)、グアニン(G)、ウラシル(U)を塩基とする核酸である。

[0003] ところで、近年、RNA干渉と呼ばれる現象が発生することが知られるようになった。RNA干渉とは、細胞内に特定の2本鎖RNAが存在することにより、特定の配列のmRNAが破壊され、遺伝子の発現が抑制される現象である。この現象は、最初、線虫の細胞を用いた実験で発見された。その後、この現象は、哺乳動物細胞でも起きることが知られるようになり、注目を集めることとなった。人為的にRNA干渉を起こすことにより、特定の遺伝子の働きを抑制することができ、その特定の遺伝子の働きを調べることができるからである。また、RNA干渉を利用することにより、特定の遺伝子の働きを抑制する効果を発揮する薬を開発できる可能性も生まれてきた。

[0004] 図1は、RNA干渉の過程の概略を示す図である。RNA干渉は、以下のようなプロセスを経て発生すると考えられている。およそ21から23塩基対の長さのsiRNA(short interfering RNA)101がマルチ・タンパク質複合体と結合し、RISC(RNA-induced silencing complex)102を形成する。RISC(102)は、そのsiRNAと相

同性を持つmRNA（103）と結合し、そのmRNA（103）を断片104、105などへ分解することにより、そのmRNA（103）が機能しなくなる。ここで、「ある塩基配列（S）と別の塩基配列（T）との間に相同性がある」とは、2つの塩基配列（S、T）が相補性を有しているか、または、不完全な相補性を有していることをいう。「相補性」とは、二つの塩基配列の全体において、AとT、GとC、AとUとの対が完全に形成されていることをいう。したがって、相同性とは、二つの塩基配列の一部に、AとT、GとC、AとU以外の対が発生していることを意味する。なお、どのような場合に、二つの塩基配列の間に相補性を有する塩基対がどれだけの存在すれば、その二つの塩基配列が相同性を有すると判断されるかについて説明すると次のようになる。すなわち、RNA干渉の場合には、80％以上、好ましくは90％以上、さらに好ましくは95％以上の場合に、相同性を有すると判断される場合が多い。また、相補性を有する塩基対の割合のみならず、相補性を有する塩基列が塩基配列中にどれだけの個数連続して現れているかを考慮にいれて、二つの塩基配列の間の相同性の有無を判断することもある。また、AとT、GとC、AとUとの3種類の相補性を有する塩基対に、GとUとの対が形成される可能性もあることが知られているので、GとUとの塩基対の存在も考慮に入れて相同性の有無を判断することもある。

[0005] したがって、RNA干渉を発生させ、目的とする遺伝子の働きを抑制するためには、siRNAの配列を設計することが重要である。すなわち、目的とする遺伝子だけに現れ、他の遺伝子の塩基配列と相同性を持たない、siRNAの配列を設計することが重要である。したがって、siRNAの配列を設計する際には、siRNAの配列に似た塩基配列を持つ遺伝子が目的とする遺伝子以外には存在しないことを確認することが必要となる。

[0006] また、近年、マイクロアレイを用いた遺伝子解析や遺伝子診断などが実施されている。「マイクロアレイ」とは、長さが15から60塩基程度のオリゴDNAをガラスなどの基板上に合成したDNAチップの一種である（例えば、非特許文献1参照。）。

[0007] 図2は、マイクロアレイを用いた遺伝子解析や遺伝子診断などの過程を例示する。ガラスなどの基板上に合成したオリゴDNAを持つマイクロアレイ201上に、蛍光色素などの標識203を付加されたDNA（202）を流すと、そのDNAと相補性あるいは相

同性を持つマイクロアレイ上のオリゴDNAとが結合(ハイブリダイズ)する(符号204)。どの場所のオリゴDNAとハイブリダイズしたかを、標識の蛍光色素による蛍光を検出することにより、DNA(202)の種類などを判定する。図2では、マイクロアレイ上に数本のオリゴDNAしか示されていないが、実際のマイクロアレイは、縦横の長さが0.5インチ程度の領域に万のオーダーでオリゴDNAが配置される。

[0008] したがって、どのような塩基配列を持つオリゴDNAをマイクロアレイに配置するかを決めることは、マイクロアレイの設計において、極めて重要な工程である。

[0009] 従来においては、似た塩基配列が存在するかどうかの検出は、BLAST(例えば非特許文献2参照。)と呼ばれるソフトウェアや、Smith－Watermanと呼ばれるアルゴリズム(例えば、非特許文献3参照。)を用いた、遺伝子情報を表わす遺伝子塩基配列を格納したデータベースの検索により行なわれている場合が多い。

非特許文献1:杉本直己著、"遺伝子化学"、19ページ、株式会社化学同人発行、2002年

非特許文献2:S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool", J. Mol. Biol. , 215, 403－410, 1990

非特許文献3:T. F. Smith, and M. S. Waterman, "Identification of common molecular subsequences", J. Mol. Biol. , 147, 195－197, 1981

発明の開示

発明が解決しようとする課題

[0010] しかしながら、BLASTを使用する方法では、似た塩基配列の存在を見落としてしまうという課題がある。BLASTは、通常、7塩基が連続して同じになる部分を用いて検索が行なわれる。このため、19塩基の塩基配列が与えられた場合、例えば、図3の×の位置で塩基の不一致あるいは不適合がある塩基配列を見つけることができず、似た塩基配列の存在を見落としてしまう。

[0011] また、Smith－Watermanアルゴリズムを用いれば、似た塩基配列の存在を正しく検出することが可能であるが、必要とされる計算量が大きく、検出までに時間がかかるという課題がある。

[0012] そこで、本発明は、似た塩基配列の存在を少ない計算量で検出することができる装置及び方法などを提供することを目的とする。

課題を解決するための手段

[0013] かかる目的を達成するために、本発明では、入力された塩基配列から所定の長さの二つの部分配列とその余の部分とを特定して、対応する塩基が適合しない塩基への置換を行なう塩基の個数であるハミング距離を、それらの部分配列とその余の部分とに分割して割り当てて、2つの部分配列のうち、それぞれの部分配列に対して、割り当てられたハミング距離の個数の塩基を適合しない塩基に置換する操作を加えて得られる塩基配列の総数が大きくないほうを選択して、検索を行なうようにする。

[0014] これにより、検索に用いる塩基配列であって塩基を置換して生成する塩基配列の数を少なくすることができ、検索のために必要となる計算量を小さくすることができ、課題が解決される。また、ハミング距離が所定の値と同じまたは所定の値未満となる、似た塩基配列の存在を見落とすこともなくなるので、課題が解決される。

発明の効果

[0015] 本発明に係る塩基配列検索装置及び塩基配列検索方法によれば、検索のために必要となる計算量を小さくすることができ、また、ハミング距離が所定の値と同じまたは所定の値未満となる、似た塩基配列の存在を見落とすことも無い。

発明を実施するための最良の形態

[0016] 以下、本発明を実施するための最良の形態について、図を用いて実施形態として説明する。なお、本発明は、これら実施の形態に何ら限定されるものではなく、その要旨を逸脱しない範囲において、種々なる態様で実施し得る。

[0017] (実施形態1:主に請求項1、9について説明する)

[0018] 本発明の実施形態1として、遺伝子情報を表わす遺伝子塩基配列を格納したデータベースにおける所定長の塩基配列の出現を検索するための索引を用いて、類似塩基配列を検索するための塩基配列検索装置として、入力された塩基配列から所定の長さの二つの部分配列とその余の部分とを特定し、対応する塩基が適合しない塩基への置換を行なう塩基の個数であるハミング距離を、それらの部分配列とその余の部分とに分割して割り当てて、2つの部分配列のうち、それぞれの部分配列に対して

、割り当てられたハミング距離の個数の塩基を適合しない塩基に置換する操作を加えて得られる塩基配列の総数が大きくないほうを選択して検索を行なう塩基配列検索装置について説明する。

[0019] ここに「対応する塩基が適合する」とは、対応する二つの塩基が、所定の二項関係を満たすことをいう。ここでいう二項関係とは、多くの場合、対を構成する塩基が合い等しいことをいう。すなわち、数学の集合論の言葉で説明すれば、二項関係が反射律のみを満たす場合に該当する。また、塩基のGとUとが結合しやすいことを考慮に入れた二項関係を用いてもよい。

[0020] なお、ここに、「所定長」とは所定の長さである。この所定の長さは、遺伝子塩基配列を格納したデータベースの検索を行なうための索引が受け付けることができる塩基配列の長さである。例えば、BLASTの場合には、通常は、所定長は7となる。また、「類似塩基配列」とは、入力された塩基配列と同じ長さで類似する塩基配列であって、前記遺伝子塩基配列に出現する塩基配列である。「類似する」とは、例えば、後で説明するように、入力された塩基配列とのハミング距離が与えられた値になることである。また、「前記遺伝子塩基配列」とは、データベースに格納されている塩基配列である。なお、索引の構造によっては、所定長は複数存在する場合がある。

[0021] このような塩基配列検索装置は、例えば、WEBブラウザに入力された塩基配列と類似の程度（例えば、ハミング距離）を受け取り、遺伝子塩基配列を格納したデータベースに対して問い合わせなどを発行して処理を行ない、結果を前記WEBブラウザに返信するサーバ装置として実現が可能である。したがって、本発明に係る塩基配列検索装置の構成要素である各部、各手段は、ハードウェア、ソフトウェア、ハードウェアとソフトウェア（プログラム）の両者、のいずれかによって構成することが可能である。たとえば、これらを実現する一例として、計算機を利用する場合には、CPU、メモリ、バス、インターフェース、周辺装置などから構成されるハードウェアと、これらのハードウェア上にて実行可能なソフトウェアを挙げることができる。

[0022] （実施形態1：構成）

図4は、本発明の実施形態1に係る塩基配列検索装置の機能ブロック図を例示する。塩基配列検索装置400は、塩基配列入力部401と、ハミング距離入力部402と、

特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407と、を有する。

[0023] 「塩基配列入力部」401は、所定長を超える長さの塩基配列を入力する。例えば、WEBブラウザから入力された塩基配列を示す情報を受信する。

[0024] 「ハミング距離入力部」402は、入力塩基配列に対して、ハミング距離を入力する。例えば、WEBブラウザから入力された数値を受信する。ここに、「入力塩基配列」とは、塩基配列入力部401に入力された塩基配列である。また、ハミング距離とは、適合しない塩基への置換の操作を行なう塩基の個数を示す値である。ハミング距離は、2つの同じ長さの塩基配列に対して定義され、対応する塩基が適合しない数であり、1つの塩基配列に対してハミング距離を指定することにより、その塩基配列のうち、ハミング距離の個数の塩基を適合しない塩基に置換して得られる塩基配列の集合が定義できる。

[0025] ハミング距離の例を挙げる。例えば、図5には、19の塩基からなる塩基配列が上下に並んでいるが、「×」で示される3箇所に対応する塩基が適合していないので、ハミング距離は3となる。図6は、ハミング距離の定義を示す。SとTがn個の塩基からなる塩基配列として、S_iをi番目の塩基とするときのSとTのハミング距離$d_H(S, T)$が定義されている。なお、Uを集合とするとき、｜U｜により集合の要素の数を表わすとし、≠は、両辺の塩基が適合しないことを意味する（例えば、同じでないことを意味する）。したがって、ハミング距離は、負でない整数となる。

[0026] 「特定部」403は、入力塩基配列の部分配列であって、所定長の長さを持ち、異なる2つの部分配列と、その余の部分と、を特定する。2つの部分配列は共通部分があってもよい。また、場合によっては、その余の部分は無くてもよい。

[0027] 図7は、特定部403により特定される2つの部分配列とその余の部分とを例示する。図7（1）においては、第一の部分配列711と第二の部分配列712とが共通部分を持たないように入力塩基配列710に存在し、入力塩基配列の両端と中央部分に、その余の部分713、714、715がある。図7（2）においては、第一の部分配列721と第二の部分配列724が入力塩基配列720の略中央に共通部分を有し、入力塩基配列720の端の部分にその余の部分723、724がある。図7（3）においては、第一の部分

配列731が入力塩基配列730の左端から伸び、また、第二の部分配列732が入力塩基配列730の右端から伸び、第一の部分配列731と第二の部分配列732は、入力塩基配列730の略中央部分で共通部分を有している。入力塩基配列がある所定長の2倍を超える場合には、図7(4)に例示されるように、入力塩基配列740の略中央がその余の部分743となる。なお、第一の部分配列と第二の部分配列の長さは所定長であるが、索引の構造によっては上述のように所定長が複数存在する場合があり、そのような場合には、第一の部分配列の長さと第二の部分配列の長さは同じであってもよいし、異なっていてもよい。

[0028] 「割当部」404は、特定部403で特定された部分配列とその余の部分とに、ハミング距離入力部402で入力されたハミング距離を分割して割り当てる。ここに、「ハミング距離を分割して割り当てる」とは、ハミング距離を負でない整数に分割して、分割で得られた整数を部分配列とその余の部分に割り振ることである。したがって、割り振られた値の総和は、ハミング距離になる。このような処理は、プログラムにより容易に実現することができる。例えば、部分配列とその余の部分の個数分のループを入れ子にするプログラムにより実現することができ、全ての割り振りを得ることができる。

[0029] 図17は、ハミング距離を分割して割り当てるプログラムの例をC言語で記述した例を示す。この例では、部分配列が数字で特定されるとする。例えば、部分配列の個数がPであれば、P、P－1、P－2、…、1でそれぞれの部分配列が特定されるとし、所定の配列のP番目、P－1番目、…、1番目の要素が、部分配列に対応しているとする。このとき、distributeHammingDistanceの第1引数には、部分配列の個数Pを渡し、第2引数には、第1引数に渡された個数の部分配列に割り当てるハミング距離Hを渡すと、P、P－1、P－2、…、1でそれぞれ特定される部分配列にハミング距離が割り当てられることになる。なお、distributeHammingDistanceの第3引数には部分配列の総数を、第4引数には、所定の配列を渡す。DistributeHammingDistanceの動作を具体的に説明すると次のようになる。すなわち、所定の配列がvecで指し示されるとすると、distributeHammingDistanceが一回呼ばれるたびに、vec[P]、vec[P－1]、vec[P－2]、…vec[1]のどれかに、部分配列に割り当てられたハミング距離が代入され、distributeHammingDistanceの再帰呼び出しがされる。例えば、

distributeHammingDistanceの或る呼び出しにおいてvec[q]に、部分配列qに割り当てられたハミング距離が代入されると、qが1でなければ、distributeHammingDistanceの第1引数をq－1にして再帰呼び出しが行なわれる。最後にvec[1]に、ハミング距離が代入されると、全ての部分配列に対するハミング距離の割り当ての一つが完成するので、vec[P]、vec[P－1]、vec[P－2]、…vec[1]の値がprintfという関数を用いて出力される。もちろん、printfによる出力を行なう代わりに、ハミング距離の割り当てをメモリに構成されるデータ構造に格納して、選択部405がそのデータ構造を参照し、後に説明されるように、部分配列の選択が行なえるようにするのは、当業者にとって容易なことである。

[0030] なお、図17のプログラムにおいて、intは、整数のデータ型を表わし、例えば、int h;は、hという変数が整数のデータ型の値をとることを意味する。また、for(S1;S2;S3){S4}は、まずS1を実行し、S2の条件が成り立つ限り、S4を実行した後にS3を実行することを繰り返すことを表わす。なお、DistributeHammingDistance は、図17に例示されているが、念のために記載すると以下のようになる。

distributeHammingDistance(int P, int H, int nSize, int* vec) { int h; if (P==1) {vec[1] = h; /* 全ての部分へのハミング距離の割り当ての一つが完成したので vec に格納されたハミング距離を出力する */ for (int i = 1; i <= nSize; i = i + 1) { printf("Part %d th: %d", i, vec[i]); /* セパレータ又はターミネータを出力する */ if (i != nSize) { /* セパレータとしてカンマを出力する */ printf(", "); } else { /* ターミネータとして改行を出力する */ printf("¥n"); } } else { for (h = 0; h <= H; h = h + 1) { vec[P] = h; distributeHammingDistance(P - 1, H - h, nSize, vec); } } }と定義される。このように、リカーシブコールを行なうことにより、部分配列とその余の部分の個数分のループを入れ子にすることができる。

[0031] 図8は、図7の(1)から(4)の場合に対応して、割当部404によるハミング距離の割り振りを説明するための図である。図8(1)において、入力塩基配列の左端の部分から(すなわち、その余の部分、第一の部分配列、その余の部分、第二の部分配列、その余の部分)、m_1、m_2、m_3、m_4、m_5という値が割り振られたとすると、m_1、m_2、m_3、m_4、m_5の和が入力ハミング距離となる。ここに、「入力ハミング距離」とは、ハミング距離

入力部402に入力されたハミング距離である。

[0032] 同様に、図8(2)においても、入力塩基配列の左端の部分から(すなわち、その余の部分、第一の部分配列の左部分、共通部分、第二の部分配列の右部分、その余の部分)、m_1、m_2、m_3、m_4、m_5が割り振られたとすると、これらの値の和が入力ハミング距離となる。

[0033] 図8(3)においては、入力塩基配列の左端の部分から(すなわち、第一の部分配列の左部分、共通部分、第二の部分配列の右部分)、m_1、m_2、m_3が割り振られたとすると、これら3つの値の和が入力ハミング距離となる。

[0034] 図8(4)においては、入力塩基配列の左端の部分から(すなわち、第一の部分配列、その余の部分、第二の部分配列)、m_1、m_2、m_3が割り振られたとすると、これら3つの値の和が入力ハミング距離となる。

[0035] 「選択部」405は、特定部403で特定された2つの部分配列のうち、前記割当部で割り当てられたハミング距離で示される個数の塩基を適合しない塩基へ置換する操作を前記部分配列に対して行なって生成される塩基配列である置換塩基配列の総数が大きくない方を選択する。この総数は、式:(適合しない塩基の数)$^{(ハミング距離)}\times{}_{(所定長)}C_{(ハミング距離)}$により計算でき、この計算結果に基づいて選択を行なう。しかし、多くの場合には、割当部404で割り当てられたハミング距離の大きくない方の部分配列を選択することになる。すなわち、図8(1)の場合には、m_2とm_4とを比較して、例えば、m_2の方が大きくなければ、第一の部分配列を選択する。逆に、m_4の方が小さければ、第二の部分配列を選択する。ただし、例えば、所定長が4である2つの部分配列に対して、一方にはハミング距離として3が、他方には4が割り当てられたとすると、一方の部分配列の置換塩基配列の総数は次のように計算される。すなわち、適合しない塩基とは、異なる塩基であるとすると、塩基の種類は4であるので、ある塩基と異なる塩基の種類の数は(4−1)となり、一方の部分配列の置換塩基配列の総数は、$(4-1)^3{}_4C_3=108$である。しかしながら、他方の部分配列の置換塩基配列の総数は、$(4-1)^4{}_4C_4=81$となり、割り当てられたハミング距離の大きな他方の置換塩基配列の数が少なくなる場合がある。したがって置換塩基配列の総数の大小とハミング距離の大小とは一致しない場合があり、注意が必要である。なお、以下では、説明を簡単に

するために、割当部404で割り当てられたハミング距離が大きくない場合が、置換塩基配列の総数が大きくないと仮定して説明を行なう。

[0036] 同様に図8(2)の場合には、m_2+m_3とm_3+m_4とを比較して、例えば、m_2+m_3の方が大きくなければ、第一の部分配列を選択する。逆に、m_3+m_4の方が小さければ、第二の部分配列を選択する。

[0037] また、図8(3)の場合には、m_1+m_2とm_2+m_3とを比較して、例えば、m_1+m_2の方が大きくなければ第一の部分配列を選択する。逆に、m_2+m_3の方が小さければ、第二の部分配列を選択する。

[0038] また、図8(4)の場合には、m_1とm_3とを比較して、例えば、m_1の方が大きくなければ、第一の部分配列を選択する。逆に、m_3の方が小さければ、第二の部分配列を選択する。

[0039] 図9は、入力ハミング距離が3であるとき、図7(4)のように部分配列とその余の部分が特定された場合の割当部404での割り振りと、選択部405による選択を示す。図9では、説明を簡略化するために、m_1、m_2、m_3の和が入力ハミング距離3と等しい場合について説明している。m_1、m_2、m_3の和が3になる組合せは、10通りあるが、選択部405で、例えば$m_1 \leqq m_3$となる選択が行なわれると、選択の結果として6通りの組み合わせが得られる。これからさらに、m_1の値の組み合せについて重複を除くと、0と1との組み合わせになる。同様のことが、第二の部分配列とm_3についても言える。結果として、m_3についても、0と1とになる。なお、$m_1=m_3$の場合は除かれるので、$m_1>m_3$となる選択が行なわれる場合の数は、$m_1 \leqq m_3$となる選択が行なわれる場合の数よりも少なくなる。このことは、後に説明する置換塩基配列生成部406と検索部407との動作により、m_1が0と1との場合について、また、m_3についても0と1との場合について、置換塩基配列生成部406で置換塩基配列を生成して検索部で索引を参照して検索を行なえば、m_1、m_2、m_3の和が3になる10通りの場合についての検索がカバーされることを意味する。

[0040] また、ハミング距離入力部に入力されたハミング距離が複数の部分に割り当てられ、$m_1 \leqq m_3$、$m_1 > m_3$のように大きくない方の選択が行なわれるので、上記のパラグラフの記述により得られるm_1、m_3の値の組み合わせは、ハミング距離入力部に入力さ

れたハミング距離未満であっても得られるものである。したがって、ハミング距離Hの場合について選択を行なうと、H未満のハミング距離がハミング距離入力部に入力された場合についての選択も行なわれる。

[0041] したがって、m_1、m_2、m_3の和が入力ハミング距離3未満である場合についても同様に処理ができる。このように、本発明では、m_1、m_2、m_3の和が入力ハミング距離が与えられた値のみならず、その値未満の場合の処理を一回の処理で実行することができる。

[0042] 図10は、同じく入力ハミング距離が3であるとき、図7(3)のように部分配列とその余の部分が特定された場合の割当部404での割り振りと、選択部405による選択を示す。図10でも、説明を簡略化するために、m_1、m_2、m_3の和が入力ハミング距離3と等しい場合について説明している。m_1、m_2、m_3の和が3になる組み合せは、同じく10通りであるが、選択部405で$m_1+m_2 \leqq m_2+m_3$となる選択が行なわれると、選択の結果として、6通りの組み合せが得られる。これらの組み合わせから第一の部分配列に割り振られたm_1+m_2の値について重複を除くと、0、1、2、3の4通りが得られる。同じことが第二の部分配列とm_2+m_3についても言える。ただし、$m_1+m_2=m_2+m_3$となる場合が除かれるので、$m_1+m_2>m_2+m_3$となる選択が行なわれる場合の数は、$m_1+m_2 \leqq m_2+m_3$となる選択がされる場合の数よりも少なくなる。この結果、m_2+m_3については、0と1との2通りが得られる。このことは、後に説明する置換塩基配列生成部406と検索部407との動作により、m_1+m_2が0、1、2、3の場合について、また、m_2+m_3についても0と1との場合について、置換塩基配列生成部406で置換塩基配列を生成して検索部で索引を参照して検索を行なえば、m_1、m_2、m_3の和が3になる10通りの場合についての検索がカバーされることを意味する。

[0043] また、上述したのと同じように、ハミング距離入力部に入力されたハミング距離が複数の部分に割り当てられ、$m_1+m_2>m_2+m_3$、$m_1+m_2 \leqq m_2+m_3$のように大きくない方の選択が行なわれるので、上記のパラグラフの記述により得られるm_1+m_2、m_2+m_3の値の組み合わせは、ハミング距離入力部に入力されたハミング距離未満であっても得られるものである。したがって、ハミング距離Hの場合について選択を行なうと、H未満のハミング距離がハミング距離入力部に入力された場合についての選択も行

なわれる。

[0044] 「置換塩基配列生成部」406は、選択部405により選択された部分配列に対して、割当部404で割り当てられたハミング距離をもつ置換塩基配列を生成する。すなわち、選択部405により選択された部分配列の塩基のうち、割当部404で割り当てられたハミング距離で示される個数の塩基を適合しない塩基に置換することを行ない、置換塩基配列を生成する。例えば、図9の場合には、第一の部分配列について、ハミング距離が0と1となる部分配列が置換塩基配列として生成される。また、第二の部分配列についても、ハミング距離が0と1となる部分配列が置換塩基配列として生成される。ハミング距離が0であれば、第一の部分配列そのものであり、ハミング距離が1であれば、第一の部分配列の塩基のうちの任意の1つを、適合しない塩基に置き換えて置換塩基配列が生成される。

[0045] 同様に図10の場合には、第一の部分配列について、ハミング距離が0、1、2、3の置換塩基配列として生成される。また、第二の部分配列についても、ハミング距離が0と1となる部分配列が置換塩基配列として生成される。この場合、入力ハミング距離が3であり、また、ハミング距離が3の置換塩基配列を生成しなければいけないのは、効率が悪いように見える。しかし、3が割り振られたのは、m_2であるので、第一の部分配列と第二の部分配列との共通部分に対して、ハミング距離が3となる置換塩基配列を生成すればよい。もし、その共通部分の長さが短ければ、ハミング距離が3となる置換塩基配列の総数は限られたものとなる。このように第一の部分配列と第二の部分配列とが共通部分を持つ場合には、共通部分とそうでない部分とに割り振られたハミング距離を考慮して、共通部分とそうでない部分に個別に置換塩基配列を生成することにより、置換塩基配列の生成の効率を上げることができる。

[0046] 置換塩基配列を生成するプログラムは容易に作成することができ、例えば、ループを入れ子にしたプログラムを作成し、外側のループにより、塩基を適合しない塩基に置換する部分配列の位置を特定し、外側のループにより特定された位置の塩基を適合しない塩基に置換することを内側のループにより行なうようにすればよい。所定長をLとし、塩基が異なるときを適合しないと定義すれば、図9の場合には、$1+3_LC_1$通りの置換塩基配列が生成される。図10の場合には、$1+3_LC_1+3^2{}_LC_2+3^3{}_LC_3$通りの

置換塩基配列が生成されるが、この生成に必要な計算量は、一般にLは入力塩基配列の長さの値よりも小さいので、入力塩基配列とハミング距離が3となる塩基配列の全てを求める計算量より小さい。

[0047] 図18は、配列Sにより長さがLの部分配列に2というハミング距離が割り当てられた場合に、その部分配列の置換塩基配列を生成するプログラムを例示する。このプログラムでは、配列の添え字は0から始まり、S[0]、S[1]、…、S[L－1]に塩基を示すA、C、G、Tのいずれかのシンボルが格納されているとする。また、例えば、foreach a1 in {A, C, G, T} {S}は、変数a1の値をA, C, G, Tに次々に変化させながら、Sを実行することを表わすとしている。図18において、for(l1＝0;l1＜L;l1＝l1＋1)とfor(l2＝0;l2＜L;l2＝l2＋1)とが上記の「外側のループ」を表わし、foreach a1 in {A, C, G, T}とforeach a2 in {A, C, G, T}とが上記の「内側のループ」を表わしている。図18にプログラムが示されているが、念のために明細書にもそのプログラムを記しておく。for (l1 = 0; l1 ＜ L; l1 = l1 + 1) { for (l2 = l1 + 1; l2 ＜ L; l2 = l2 + 1) { foreach a1 in {A, C, G, T} { if (S[l1] != a1) { foreach a2 in {A, C, G, T} { if (S[l2] != a2) { S のl1番目の塩基をa1に置換し、Sのl2番目の塩基をa2に置換して得られる置換塩基配列を生成; } } } } }。

[0048] 「検索部」407は、置換塩基配列生成部で生成された置換塩基配列をキーとして前記索引を用いて検索を行なう。多くの場合、索引はハッシュの手法を用いて実現されている。「前記索引」とは、所定長の塩基配列の、遺伝子配列を格納したデータベースにおける出現を検索するための索引である。このような索引による検索により、一般には、置換塩基配列が出現する位置情報（例えば、置換塩基配列の端の塩基が、DNAの5´端から何番目の位置の塩基になるかを示す情報）が得られる。

[0049] もし、塩基配列検索装置が遺伝子塩基配列を格納したデータベースを備えていれば、検索部407は、そのデータベースに対して問い合わせを行なう。また、他のサーバであって、そのようなデータベースを備えているサーバがあれば、検索部407はそのサーバに問い合わせを送信して、結果を受信するようになっていてもよい。

[0050] （実施形態1：処理の流れ）

図11は、本実施形態に係る図4の塩基配列検索装置の処理の流れ図を例示する

。ステップS1101において、塩基配列入力部401などにより、塩基配列を入力する(塩基配列入力ステップ)。ステップS1102において、ハミング距離入力部402などにより、ハミング距離を入力する(ハミング距離入力ステップ)。ステップS1103において、特定部403などにより、2つの部分配列とその余の部分とを特定する(特定ステップ)。ステップS1104において、割当部404などにより、入力されたハミング距離を分割して割り当てる(割当ステップ)。ステップS1105において、選択部405などにより、割当ステップで割り当てられたハミング距離を有する置換塩基配列の総数の大きくない方の部分配列を重複が発生しないように選択する(選択ステップ)。ステップS1106において、置換塩基配列生成部406などにより、置換塩基配列を生成する(置換塩基配列生成ステップ)。ステップS1107において、検索部407などにより、検索を行なう(検索ステップ)。

[0051] したがって、塩基配列検索装置は、塩基配列入力ステップ、ハミング距離入力ステップ、特定ステップ、割当ステップ、選択ステップ、置換塩基配列生成ステップ、検索ステップを含む塩基配列検索方法を使用するための装置とみなすこともできる。

[0052] なお、図11に例示された流れ図は一例であり、ステップS1101で入力された塩基配列の一つについて、ステップS1102で入力されるべきハミング距離を0、1、2、3、4などと変化させながら、その他のステップを繰り返し実行してもよい。また、ステップS1101を行なった後でステップS1103を行ない、ステップS1102を実行して、その他のステップを実行するようになっていてもよい。入力するハミング距離を0、1、2、3、4などと変化させながら、ステップS1101からステップS1104までを実行した後に、まとめてステップS1105以下を実行するようにしてもよい。このようにすることにより、同じ部分配列を用いた検索を再度繰り返すことなく、効率よく計算を進めることができる。

[0053] (実施形態1:主な効果)

本実施形態により、検索のために必要となる計算量を小さくすることができ、また、ハミング距離が所定の値、もしくはそれ以下、もしくは任意の値の組み合わせとなる、似た塩基配列を漏れなく検索することができる。

[0054] なお、図4の機能ブロック図により表わされる塩基配列検索装置の構成は、ハードウェアとしては、任意の計算機のCPU、メモリ、その他のLSIなどにより実現することが

できる。また、ソフトウェアとしては、メモリにロードされたプログラムなどにより実現することができる。また、ハードウェアとソフトウェアとの連携により実現することもできる。特にソフトウェアが用いられて実現される場合には、そのようなソフトウェアを構成するプログラムを、各種の媒体に記録しておき、必要に応じて塩基配列検索装置を実現するための計算機に機械的に読み取られるようにすることができる。ここで、「媒体」とは、フレキシブルディスク、光磁気ディスク、ROM、EPROM、EEPROM、CD－ROM、MO、DVD、フラッシュディスク等の任意の「可搬用の物理媒体」や、各種計算機システムに内蔵されるROM、RAM、HD等の任意の「固定用の物理媒体」、あるいはLAN、WAN、インターネットに代表されるネットワークを介してプログラムを送信する場合の通信回線や搬送波のように短期にプログラムを保持する「通信媒体」を含むものとする。なお、ここにいう計算機とは、メインフレーム計算機に限定されることはなく、ワークステーションやパーソナルコンピュータなどの情報処理装置であってもよい。また、そのような情報処理装置には、プリンタやスキャナなどの周辺装置がさらに接続されていてもよい。

[0055] また、「プログラム」とは、任意の言語や記述方法にて記述されたデータ処理方法であり、ソースコードやバイナリコート等の形式を問わない。なお、「プログラム」は必ずしも単一的に構成されるものに限られず、複数のモジュールやライブラリとして分散構成されるものや、オペレーティングシステムに代表される別個のプログラムと協同してその機能を達成するものをも含む。なお、塩基配列検索装置において媒体を読み取るための具体的な構成、読み取り手段、あるいは、読み取り後のインストール手順等は、周知の構成や手順を用いることができる。例えば、本実施形態に係る塩基配列検索装置の、塩基配列入力部401と、ハミング距離入力部402と、特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407とは、それぞれプログラムを構成するモジュールとして実現することができる。そのようなモジュールは、当然、計算機のCPUにより制御を受けることとなる。

[0056] 本明細書では図示を省略しているが、塩基配列検索装置は、遺伝子の塩基配列情報等に関する外部データベースの検索などを行なうための外部プログラム等を提供する外部システムに、インターネット等の通信網を介して通信可能に接続された構

成であってもよい。かかる構成により、外部プログラムを実行するウェブサイトが提供される。外部システムは、WEBサーバやASPサーバ等として構成されてもよい。例えば、塩基配列検索装置が外部システムに通信可能に接続されてもよい。通信網の構成は特には限定されないが、例えば、ルータ等の通信装置や専用線等の有線又は無線の通信回線により構成される。

[0057] （実施形態2：主に請求項2について説明する）

[0058] 図12は、本発明の実施形態2に係る塩基配列検索装置の機能ブロック図を例示する。塩基配列検索装置1200は、塩基配列入力部401と、ハミング距離入力部402と、特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407と、を有し、特定部403は、第一特定手段1201を有している。したがって、本実施形態に係る塩基配列検索装置は、実施形態1に係る塩基配列検索装置の特定部が第一特定手段を有した構成となっている。

[0059] 「第一特定手段」1201は、塩基配列入力部で入力された塩基配列の塩基数が前記所定長の2倍以下または2倍未満であれば、前記2つの部分配列のうち一方の部分配列の端を前記入力塩基配列の他方の端と一致させ、その余の部分が生じず特定されないことにする。その余の部分が生じず特定されないことにより、割当部では、その余の部分にハミング距離を割り当てることはしないこととなる。

[0060] すなわち、第一特定手段は、図7（3）のように第一の部分配列と第二の部分配列とを特定する。したがって、このような場合は実施形態1について既に説明されているので、以後の説明は省略する。

[0061] （実施形態3：主に請求項3について説明する）

[0062] 図13は、本発明の実施形態3に係る塩基配列検索装置の機能ブロック図を例示する。塩基配列検索装置1300は、塩基配列入力部401と、ハミング距離入力部402と、特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407と、を有し、特定部403は、第二特定手段1301を有している。また、特定部403は、実施形態2で説明した第一特定手段を有していてもよい。したがって、本実施形態に係る塩基配列検索装置は、実施形態1または2に係る塩基配列検索装置の特定部が第二特定手段を有した構成となっている。

[0063] 「第二特定手段」1301は、塩基配列入力部で入力された塩基配列の塩基数が前記所定長の2倍より大であれば、前記2つの部分配列が重ならないことにして、前記2つの部分配列を特定する。この場合、その余の部分が一つになるようにしてもよいし、2つになるようにしてもよい。例えば、2つの部分配列が入力塩基配列の左右の端に配置されるように特定したり、2つの部分配列が連接されるように入力塩基配列を特定したりする。

[0064] すなわち、第二特定手段は、図7(4)のように第一の部分配列と第二の部分配列とを特定する。したがって、このような場合は実施形態1について既に説明されているので、以後の説明は省略する。

[0065] (実施形態4:主に請求項4について説明する)

[0066] 本発明の実施形態4として、検索部での検索結果に基づいて、類似塩基配列の候補を取得して、入力塩基配列とのハミング距離を判定する塩基配列検索装置について説明する。

[0067] (実施形態4:構成)

図14は、本発明の実施形態4に係る塩基配列検索装置の機能ブロック図を例示する。塩基配列検索装置1400は、塩基配列入力部401と、ハミング距離入力部402と、特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407と、類似候補塩基配列取得部1401と、判定部1402と、を有している。また、特定部403は、実施形態2で説明した第一特定手段と実施形態3で説明した第二特定手段とのいずれか一方または両方を有していてもよい。したがって、本実施形態に係る塩基配列検索装置は、実施形態1から3のいずれか一の塩基配列検索装置が類似候補塩基配列取得部1401と、判定部1402と、を有している構成となっている。

[0068] 「類似候補塩基配列取得部」1401は、検索部407での検索結果に基づいて、類似候補塩基配列を取得する。「類似候補塩基配列」とは、置換塩基配列を含んで遺伝子塩基配列に現れる塩基配列である。より具体的に説明すると、例えば、第一の部分配列の置換塩基配列により検索が行なわれ、置換塩基配列の端の塩基の位置が判明したとすると、第一の部分配列が入力塩基配列に占める位置関係を勘案して得られる塩基配列で、入力塩基配列と同じ長さの遺伝子塩基配列を取得する。すな

わち、検索で得られる位置が第一の部分配列の左端の塩基の位置であるとすると、第一の部分配列の左側のその余の部分の長さ(もし、そのようなその余の部分がなければ0とする)だけ左の位置から、入力塩基配列と同じ長さの遺伝子塩基配列を取得する。第二の部分配列の置換塩基配列について検索が行なわれた場合も同様に、第二の部分配列の右側のその余の部分の長さだけ右の位置から左に向かって入力塩基配列と同じ長さの遺伝子塩基配列を取得する。この取得は、データベースを検索することにより行なわれる。もし、塩基配列検索装置が、そのようなデータベースを備えていれば、そのデータベースから取得を行ない、別のサーバにそのようなデータベースが備えられていればそのサーバに取得の要求を送信して、類似候補塩基配列を得る。

[0069] 「判定部」1402は、類似候補塩基配列取得部で取得された類似候補塩基配列と、前記入力塩基配列と、のハミング距離が、ハミング距離入力部402に入力されたハミング距離、もしくはそれ以下、または入力されたハミング距離の組に一致するかどうかを判定する。この判定は、入力塩基配列と類似候補塩基配列との端の塩基から順に比較を行なうことにより行なうことができる。

[0070] 本実施形態に係る塩基配列検索装置の処理の流れ図は、図11に例示された流れ図のステップS1107の後に、類似候補塩基配列を取得するステップと、類似候補塩基配列と入力塩基配列とのハミング距離が、入力ハミング距離に等しいかどうかを判定するステップと、を実行する流れ図となる。

[0071] (実施形態4:主な効果)

本実施形態によれば、入力塩基配列に類似する塩基配列を取得することができ、例えば、siRNAにより不活性化する目的の遺伝子以外に不活性化される可能性のある遺伝子の情報を得ることが可能となる。

[0072] (実施形態5:主に請求項5について説明する)

[0073] 本発明の実施形態5として、不適合となる塩基の組合せを指定することができる塩基配列検索装置について説明する。

[0074] (実施形態5:構成)

図15は、本発明の実施形態5に係る塩基配列検索装置の機能ブロック図を例示す

る。塩基配列検索装置1500は、塩基配列入力部401と、ハミング距離入力部402と、特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407と、類似候補塩基配列取得部1401と、判定部1402と、不適合塩基組入力部1501と、を有している。したがって、本実施形態に係る塩基配列検索装置は、実施形態4に係る塩基配列検索装置が不適合塩基組入力部1501を有している構成となっている。

[0075] 「不適合塩基組入力部」1501は、適合しない塩基の組を指定する。例えば適合しないと判断するべき塩基のペアを示すテキスト情報を入力する。あるいは、適合と判断するべき塩基のペア(例えば、GとU)を入力することにより、間接的に適合しないと判断するべき塩基の組が指定されるようになっていてもよい。

[0076] 本実施形態においては、不適合塩基組入力部1501に入力された塩基の組に基づいて検索部で検索が行なわれ、また、ハミング距離が求められる。例えば、不適合塩基組入力部1501により入力された塩基の組に基づいて、置換塩基配列が置換塩基配列生成部406で生成され、検索部407では検索のためのデータベースが選択され、判定部1402でハミング距離が求められる。

[0077] (実施形態5:主な効果)

本実施形態によれば、例えば、GとUのように弱いながらも結合する可能性のある塩基の組合せを考慮することができ、より正確な塩基配列の設計を行なうことが可能となる。

[0078] (実施形態6:主に請求項6について説明する)

[0079] 本発明の実施形態6として、入力塩基配列と類似塩基配列との塩基の適合の分布を指定することができる塩基配列検索装置について説明する。

[0080] (実施形態6:構成)

図16は、本発明の実施形態6に係る塩基配列検索装置の機能ブロック図を例示する。塩基配列検索装置1600は、塩基配列入力部401と、ハミング距離入力部402と、特定部403と、割当部404と、選択部405と、置換塩基配列生成部406と、検索部407と、類似候補塩基配列取得部1401と、判定部1402と、適合分布入力部1601と、を有しており、判定部1402は、判定手段1602を有している。また、塩基配列検

索装置1600は、実施形態5にて説明した不適合塩基組入力部を有していてもよい。したがって、本実施形態に係る塩基配列検索装置は、実施形態4または5に係る塩基配列検索装置が、適合分布入力部1601を有し、判定部1402は、判定手段1602を有している構成となっている。

[0081] 「適合分布入力部」1601は、塩基配列入力部401に入力された塩基配列と類似塩基配列との対応する塩基の適合の分布を表わす分布情報を入力する。分布情報の例としては、5´端側の方に塩基の不適合の発生が少ない、あるいは、多い、塩基の不適合がほぼ等間隔で発生していることを示す情報がある。分布情報は、例えば、塩基の適合の分布を判定するプログラムであってもよい。あるいは、あらかじめ塩基の適合の分布の類型をいくつか決めておき、それらを選択するための情報であってもよい。

[0082] 「分布判定手段」1602は、適合分布入力部1602で入力された分布情報が満たされているかどうかを判定する。

[0083] 判定部1402は、例えば、類似塩基配列とともに、分布判定手段での判定の結果を表示するようになっていてもよい。

[0084] (実施形態6:主な効果)

本実施形態により、より正確な塩基配列の設計を行なうことが可能となる。

[0085] (実施形態7:主に請求項7について説明する)

[0086] 本発明の実施形態7に係る塩基配列検索装置は、実施形態6に係る塩基配列検索装置において、適合分布入力部1601で入力される分布情報を、塩基配列と類似塩基配列との対応する塩基が連続して適合する長さの下限としたものである。

[0087] 2つの塩基配列において、対応する塩基に不適合となるものがあっても、対応する塩基が連続して適合していると、結合(ハイブリダイズ)してしまう場合がある。本実施形態においては、塩基が連続して適合する長さの下限を指定することにより、結合してしまう可能性のある類似塩基配列を検出するようにしたものである。

[0088] (実施形態8:主に請求項8について説明する)

[0089] 本発明の実施形態8は、実施形態1から7のいずれか一の実施形態において、塩基配列入力部に入力される塩基配列の長さを15から60まで、望ましくは15から25ま

でとし、所定長を11から14とした実施形態である。

[0090] 塩基配列入力部に入力される塩基配列の長さを15から60まで、望ましくは15から25までとすることにより、本実施形態に係る塩基配列検索装置をsiRNAの設計に適したものとすることができる。また、発明者がベンチマークテストに用いたデータベースでは、入力塩基配列の長さが19または20のときには、所定長を11から14とした場合が、最も高速に検索が行なえた。これは、所定長が小さいと、類似候補塩基配列の候補の数が多くなり、一方、所定長を大きくすると、置換塩基配列生成部での置換塩基配列の生成に計算量が必要となるとともに、索引を構成するハッシュテーブルに対して問い合わせを行なった際のミスヒットが増加する、すなわち、もともとのデータベース中に存在しない配列を問い合わせすることになる場合が増え、計算量が増加するためためであり、その中間点が、所定長が11から14である場合と考えられる。また、塩基配列入力部に入力される塩基配列の長さは、19または20に限定されることなく、15から60までは実用的に検索を行なうことができることが確認できた。なお、61以上になると急激にパフォーマンスの低下などが発生し実用に堪えなくなるというわけではなく、入力される塩基配列の長さが大きくなるにつれて徐々にパフォーマンスが低下することが確認された。したがって、60程度の長さのオリゴDNAの配列の決定にも本発明は使用することができることが確認できている。

[0091] (実施形態9:主に請求項10、11について説明する)

[0092] 以上、データベースに格納された遺伝子塩基配列に対する検索について述べたが、本発明の技術は、遺伝子塩基配列に限らず、一般の文字列検索などに応用することができる。すなわち、遺伝子塩基配列は、4つの塩基が一次元に配列したものであるので、それぞれの塩基を、文字列を構成するアルファベットとみなすことにより、遺伝子塩基配列を文字列とみなすことができる。また、上記の説明から判明するように、塩基の数が4である点は、本発明の技術を一般の文字列に対して適用する制限とはならない。

[0093] したがって、本発明の技術により、データベースに蓄積された文字列から、入力された文字列に類似する文字列を検索することが可能となる。ここに「類似する」とは、入力された文字列から所定のハミング距離となる文字列、または入力された文字列

から所定のハミング距離未満となる文字列を意味する。

[0094] したがって、次の文字列検索装置が提供される。すなわち、アルファベットが一次元に配列した文字列を格納したデータベースを検索するための索引であり、所定の長さである所定長の文字列が前記データベースに格納された文字列の中に出現する位置を検索するための索引、を用いて、入力される文字列と同じ長さで類似する文字列であり前記前記データベースに格納された文字列に出現する文字列である類似文字列を検索するための文字列検索装置であって、前記所定長を超える長さの文字列を入力する文字列入力部と、前記文字列入力部に入力された文字列である入力文字列に対して、適合しないアルファベットへの置換の操作を行なうアルファベットの個数を示すハミング距離を入力するハミング距離入力部と、前記入力文字列の部分文字列であって、前記所定長の長さを持ち異なる2つの部分文字列と、その余の部分と、を特定する特定部と、前記特定部で特定された部分文字列とその余の部分とに、前記ハミング距離入力部で入力されたハミング距離を分割して割り当てる割当部と、前記特定部で特定された2つの部分文字列のうち、前記割当部で割り当てられたハミング距離で示される個数のアルファベットを適合しないアルファベットへ置換する操作を前記部分文字列に対して行なって生成される文字列である置換文字列の総数が大きくない方を選択する選択部と、前記選択部により選択された部分文字列に対して、前記割当部で割り当てられたハミング距離をもつ置換文字列を生成する置換文字列生成部と、前記置換文字列生成部で生成された置換文字列を検索キーとして前記索引を用いて検索を行なう検索部と、を有する文字列検索装置を提供することが可能となる。

[0095] また、文字列のアルファベットをペプチドとすることにより、本発明の技術をペプチド配列の類似検索、すなわち、入力されたペプチド配列に類似のペプチドを検索することにも使用することができる。

[0096] （実施形態10：主に請求項12について説明する）

[0097] 本発明の実施形態10として、実施形態1から8のいずれかの塩基配列検索装置について、リピート配列の検索について改良を行なった実施形態について説明する。

[0098] （実施形態10：構成）

図19は、本発明の実施形態10に係る塩基配列検索装置の機能ブロック図を例示する。本実施形態に係る塩基配列検索装置は、実施形態1から8のいずれかの塩基配列検索装置が、リピート配列蓄積部1901と、リピート配列情報蓄積部1902と、を有し、検索部407が、リピート配列判定手段1903と、リピート配列検索手段1904と、を有する構成となっている。図19は、実施形態1に係る塩基配列検索装置が、これらの部、手段を有する場合の機能ブロック図である。

[0099] 「リピート配列蓄積部」1901は、遺伝子塩基配列中に繰り返して出現する前記所定長の塩基配列を蓄積する。「前記所定長」とは、塩基配列検索装置が用いる索引によって定まる値であり、塩基配列が遺伝子塩基配列のどの位置に現れるかをその索引により検索できるような塩基配列の長さである。

[0100] 遺伝子塩基配列の中に同じ塩基配列が複数回出現することが知られており、塩基配列によっては、その塩基配列の種類は少ないが、膨大な回数にのぼって遺伝子塩基配列に出現することが知られている。もし、置換塩基配列生成部406で生成される置換塩基配列がこのような膨大な回数にのぼって遺伝子塩基配列に出現すると、実施形態1から8の塩基配列検索装置の行なう処理の効率を低下させる。そこで、本実施形態では、置換塩基配列生成部406で生成される置換塩基配列が、遺伝子塩基配列中に繰り返して出現する場合を特別に扱うことにする。このために、まず、遺伝子塩基配列中に繰り返して出現する塩基配列をリピート配列蓄積部1901に蓄積する。

[0101] 図20は、遺伝子塩基配列中に繰り返して出現する塩基配列を表に格納した状態を例示する。遺伝子塩基配列中に繰り返して出現する塩基配列を一意に識別する識別子とその塩基配列を同じ行に格納することにより、識別子と塩基配列を関連づけて表に格納している。

[0102] 「リピート配列情報蓄積部」1902は、リピート配列情報を蓄積する。リピート配列情報とは、リピート配列蓄積部1901に蓄積された塩基配列に、その塩基配列の遺伝子配列中における出現位置を関連付けた情報である。

[0103] 図21は、リピート配列情報を蓄積するための表を例示する。この表では、図20の表で使用されている識別子と、塩基配列が遺伝子塩基配列の中に出現する位置と、を

同じ行に格納することにより、関連づけを行なっている。「リピート配列識別子」という名前の列には、識別子が格納され、「出現位置」という名前の列には、塩基配列が遺伝子塩基配列の中に出現する位置が格納されている。

[0104] 「リピート配列判定手段」1903は、置換塩基配列生成部406で生成された置換塩基配列が、リピート配列蓄積部1901に蓄積されているかどうかを判定する。例えば、図20の表の「リピート配列」という名前の列に、置換塩基配列が格納されているかどうかを調べる。この処理は、キーとして「リピート配列」という名前の列に格納されている塩基配列を持ち、バリューとして「リピート配列識別子」という名前の列に格納されている識別子を持つ索引(例えば、B^+木により構成されるもの)を用いることにより、高速に行なうことができる。なお、リピート配列判定手段1903により、リピート配列蓄積部1901に蓄積されていると判定される塩基配列をリピート配列と呼ぶことにする。

[0105] 「リピート配列検索手段」1904は、リピート配列判定手段1903にて、置換塩基配列がリピート配列蓄積部1901に蓄積されていると判定された場合には、リピート配列情報蓄積部1902に蓄積されたリピート配列情報に基づいて検索を行なう。例えば、図20の表よりリピート配列識別子という列に格納されている識別子を得て、図21の表より出現位置を求め、遺伝子塩基配列におけるその出現位置の前後の塩基配列を取得して、その塩基配列が入力塩基配列と所定のハミング距離以下であるかどうかの判断を行なうなどして検索を行なう。

[0106] (実施形態10:処理の流れ)

図22は、本実施形態に係る図19の塩基配列検索装置の検索部での処理の流れを説明するフローチャートを例示する。ステップS2201において、リピート配列判定手段により、置換塩基配列がリピート配列であるかどうかを判定する。もし、リピート配列である場合(すなわち、ステップS2201においてYESに分岐する場合)ならば、処理をステップS2202へ進め、リピート配列検索手段1904により、リピート配列情報に基づいて検索を行なう。もし、リピート配列でない場合(すなわち、ステップS2201においてNOへ分岐する場合)ならば、ステップS2203へ処理を進め、実施形態1ないし8による類似塩基配列の検索を行なう。また、リピート配列である場合ならば検索を行なわず、リピート配列でないと判断された場合のみ検索することも可能である。

[0107] （実施形態10：主な効果）

本実施形態では、置換塩基配列がリピート配列である場合には、リピート配列用の検索処理を行なうことにより、リピート配列による検索スピードの低下を防止することができる。

[0108] （実施形態11：主に請求項13について説明する）

[0109] 本発明の実施形態11として、類似塩基配列の検索結果を蓄積する塩基配列検索装置について説明する。

[0110] （実施形態11：構成）

図23は、本発明の実施形態11に係る塩基配列検索装置の機能ブロック図を例示する。本実施形態に係る塩基配列検索装置は、実施形態4から7のいずれかの塩基配列検索装置が、類似塩基配列蓄積部2301を有する構成となっている。図23は、実施形態4に係る塩基配列検索装置が、類似塩基配列蓄積部2301を有する場合の機能ブロック図である。

[0111] 「類似塩基配列蓄積部」2301は、判定部1402にて、入力塩基配列と、類似候補塩基配列取得部1401により取得された類似塩基配列と、のハミング距離がハミング距離入力部402に入力されたハミング距離以下であると判定された場合、(1)その入力塩基配列と、(2)その入力塩基配列とその類似塩基配列とのハミング距離と、(3)その類似塩基配列と、を関連付けて蓄積する。

[0112] 図24は、(1)入力塩基配列と、(2)その入力塩基配列とその類似塩基配列とのハミング距離と、(3)その類似塩基配列と、を関連付けて蓄積するための表の構造を例示する。「入力塩基配列」、「ハミング距離」、「類似塩基配列」という名前のそれぞれの列に、(1)入力塩基配列と、(2)その入力塩基配列とその類似塩基配列とのハミング距離と、(3)その類似塩基配列と、が格納される。

[0113] （実施形態11：処理の流れ）

図25は、本実施形態に係る塩基配列検索装置の判定部と類似塩基配列蓄積部との処理の流れを説明するフローチャートを例示する。ステップS2501において、判定部により、入力塩基配列と類似塩基配列とのハミング距離が入力されたハミング距離であるかどうかを判定する。もし、そうであれば、ステップS2501のYESの枝へ分岐し

、ステップS2502において、類似塩基配列蓄積部2301に、(1)入力塩基配列と、(2)ハミング距離と、(3)類似塩基配列と、を関連付けて蓄積する。ステップS2501でNOの枝へ分岐する場合には、ステップS2502は実行しない。

[0114] (実施形態11:主な効果)

本実施形態では、塩基配列検索装置の検索結果が類似塩基配列蓄積部2301に蓄積されるので、もし、既に検索対象と同じ入力塩基配列と同じハミング距離とに対して検索が行なわれているかどうかを、類似塩基配列蓄積部2301に蓄積された情報を検索して判断することにより、類似塩基配列の検索を効率よく行なうことができる。本実施形態に係る塩基配列検索装置は、例えば、インターネットなどにより検索のサービスを多数の人に提供する場合に特に有用である。例えば、第一の人が検索を行ないその後、第二の人が同じ検索を行なった場合、第二の人には、第一の人に対して提供した検索の結果を流用することにより、応答時間の短縮や、塩基配列検索装置の負荷の低減を行なうことができる。

[0115] (実施形態12:主に請求項14について説明する)

[0116] 本発明の実施形態12として、会合率を計算する塩基配列検索装置について説明する。ここに「会合率」とは、2種類の塩基配列を液体の中などの流動性のある環境下に置いた場合、どれだけの割合でその2種類の塩基配列が結合するかを示す値である。このような値は、塩基配列より物理化学的な計算を行なうことにより計算することができる。例えば、上記の非特許文献1として挙げた文献にその計算方法が開示されている。

[0117] (実施形態12:構成)

図26は、本発明の実施形態12に係る塩基配列検索装置の機能ブロック図を例示する。本実施形態に係る塩基配列検索装置は、実施形態4から7のいずれかの塩基配列検索装置が、会合率計算部2601を有する構成となっている。図26は、実施形態4に係る塩基配列検索装置が、会合率計算部2601を有する場合の機能ブロック図である。

[0118] 「会合率計算部」2601は、類似候補塩基配列取得部1401により取得された類似候補塩基配列と塩基配列入力部401により入力された入力塩基配列とのハミング距

離がハミング距離入力部402に入力されたハミング距離以下であると判定された場合に、(1)塩基配列入力部401により入力された入力塩基配列と(2)類似候補塩基配列取得部1401で取得された類似候補塩基配列との会合率を計算する。例えば、液体の温度、pHなどの条件を設定しておき、その条件での会合率を物理化学的に計算する。なお、会合率を計算する場合には、入力塩基配列を構成する塩基または類似候補塩基配列を構成する塩基を相補的な塩基に置換する。

[0119] (実施形態12:主な効果)

本発明の塩基配列検索装置では、入力塩基配列とハミング距離が所定の値以下の塩基配列を効率よく検索することができ、しかも、実際にウェット実験を行なった場合にどれだけの会合率となるかを得ることができ、実験結果やRNA干渉を用いた薬の効果の予測などを行なうことができる。

[0120] (実施形態13:主に請求項15について説明する)

[0121] 本発明の実施形態13として、ウェット実験などでコントロールとして用いることができる塩基配列を検索する装置について説明する。

[0122] (実施形態13:構成)

図27は、本発明の実施形態13に係る無効果塩基配列生成装置の機能ブロック図を例示する。無効果塩基配列生成装置2700は、塩基配列取得部2701と、無効果候補置換塩基配列生成部2702と、無効果候補置換塩基配列入力部2703と、第二ハミング距離入力部2704と、選択部2705と、を有する。

[0123] 「塩基配列取得部」2701は、前記所定長を超える長さの塩基配列を取得する。「前記所定長」とは、実施形態10で説明したように、実施形態4から7のいずれかに係る塩基配列検索装置が用いる索引によって定まる値であり、塩基配列が遺伝子塩基配列のどの位置に現れるかをその索引により検索できるような塩基配列の長さである。塩基配列取得部は、例えば、通信網を介してクライアント装置と接続され、そのクライアント装置で動作するWEBブラウザなどに入力された塩基配列を取得する。塩基配列取得部2701が取得する塩基配列は、例えば、目的とするmRNAの機能をさせないことが判明した塩基配列である。

[0124] 「無効果候補置換塩基配列生成部」2702は、無効果候補置換塩基配列を生成す

る。「無効果候補置換塩基配列」とは、塩基配列取得部で取得された塩基配列の塩基のうち、所定の個数の塩基を置換して得られる塩基配列である。例えば、塩基配列の長さが21であり、所定の個数が3であれば、$(4-1)^3{}_{21}C_3$の個数の無効果候補置換塩基配列を生成する(「4－1」の4は、塩基の種類が4であることを示す)。また、全ての無効果候補置換塩基配列するのではなく、特別な知見に基づいて目的とするmRNAの塩基配列と会合率が低くなると予測される塩基配列を生成するようにしてもよい。また、出現回数の少ない配列を用いて無効化候補置換塩基配列を生成するようにしてもよい。

[0125] 「無効果候補置換塩基配列入力部」2703は、無効果候補置換塩基配列生成部2702で生成された無効果候補置換塩基配列を実施形態12に係る塩基配列検索装置2706に入力する。例えば、無効果塩基配列生成装置と実施形態12に係る塩基配列検索装置とがLANなどで接続されていれば、実施形態12に係る塩基配列検索装置へ向けて無効果候補置換塩基配列を表わす情報を送信する。

[0126] 「第二ハミング距離入力部」2704は、無効果候補置換塩基配列入力部2703が無効果候補置換塩基配列を入力した塩基配列検索装置2706に所定のハミング距離を入力する。例えば、無効果候補置換塩基配列入力部2703が無効果候補置換塩基配列を入力するときに所定のハミング距離を入力する。

[0127] 「選択部」2705は、無効果候補置換塩基配列入力部の入力と第二ハミング距離入力部2704の入力とにより塩基配列検索装置2706より得られた会合率の低い塩基配列を選択する。例えば、ある無効果候補置換塩基配列とそれに類似する類似塩基配列との会合率が50%であり、別の無効果候補置換塩基配列とそれに類似する類似塩基配列との会合率が10%であれば、後者の無効果候補置換塩基配列を選択し、効果の無い塩基配列として無効果塩基配列生成装置の利用者に表示などする。

[0128] (実施形態13:処理の流れ)

図28は、本実施形態に係る無効果塩基配列生成装置の処理の流れを説明するフローチャートを例示する。ステップS2801において、塩基配列を、塩基配列取得部2701により取得する。ステップS2802において、無効果候補置換塩基配列を、無効果候補置換塩基配列生成部2702により生成する。ステップS2803において、塩基

配列検索装置2706に、無効果候補置換塩基配列と所定のハミング距離を入力する。ステップS2803は、個々の無効果候補置換塩基配列に対して一回ずつ行なわれ、個々の無効果候補置換塩基配列に対して会合率が取得される。ステップS2804においては、会合率の低い無効果候補置換塩基配列を、選択部2705により選択する。

[0129] （実施形態13：主な効果）

本実施形態により、与えられた塩基配列に似た塩基配列であって、会合率の低いものを選択することができる。選択により得られた塩基配列は、効果のない塩基配列と推定されるので、ウェット実験におけるコントロールなどとして用いることができる。

[0130] （実施形態14：主に請求項16について説明する）

[0131] 本発明の実施形態14として、本発明の塩基配列検索装置を用いた塩基配列のアラインメントを行なう装置について説明する。

[0132] 図29は、本発明の実施形態14における装置による処理の概要を説明するための図である。遺伝子塩基配列2901があるとして、この配列のどの部分に、塩基配列2902と似た塩基配列が存在するかを知りたいとする。この場合において、塩基配列2902の部分配列2903を得る。部分配列2903の長さは、本発明の塩基配列検索装置に適した長さであり、望ましくは15から25である。そして、本発明の塩基配列検索装置を用いて、部分配列2903の類似塩基配列2904を遺伝子塩基配列2901の中に見つける。その後、部分配列2903と類似塩基配列2904との前後の塩基の配列を、ダイナミックプログラミングなどによる従来知られている手法を用いて、比較する。このような操作により、遺伝子塩基配列2901のどの部分に塩基配列2902と似た塩基配列が存在するかを効率良く知ることができる。

[0133] （実施形態14：構成）

図30は、本発明の実施形態14に係る塩基配列アラインメント装置の機能ブロック図を例示する。塩基配列アラインメント装置3000は、第二塩基配列取得部3001と、部分塩基配列選択部3002と、部分塩基配列入力部3003と、第三ハミング距離入力部3004と、アラインメント部3005と、を有する。

[0134] 「第二塩基配列取得部」3001は、前記所定の長さを超える塩基配列を取得する。

[0135] 「部分塩基配列選択部」3002は、第二塩基配列取得部3001で取得された塩基配列の一部分である部分塩基配列を選択する。例えば、第二塩基配列取得部3001で取得された塩基配列から長さが15から25の長さの塩基配列を選択する。取得される部分塩基配列は、実施形態12で説明したリピート配列にならないのが望ましい。なぜなら、アラインメントの候補が多数発見されてしまい後に説明するステップS3104を多くの回数実行しなければいけなくなるからである。そのため、実施形態12のように、リピート配列蓄積部が塩基配列アラインメント装置に備わっており、そのリピート配列蓄積部に蓄積された内容を参照して、部分塩基配列が取得されるようになっていてもよい。

[0136] 「部分塩基配列入力部」3003は、部分塩基配列選択部で選択された部分塩基配列を実施形態4から8のいずれかに係る塩基配列検索装置3006に入力する。

[0137] 「第三ハミング距離入力部」3004は、所定のハミング距離を部分塩基配列入力部が部分塩基配列を入力した塩基配列検索装置3006に入力する。部分塩基配列入力部3003と第三ハミング距離入力部3004とによるそれぞれの入力により、部分塩基配列の類似塩基配列が求まり、遺伝子塩基配列中での位置が求まる。

[0138] 「アラインメント部」3005は、部分塩基配列入力部3003による入力と第三ハミング距離入力部3004による入力とが行われることによって塩基配列検索装置3006より得られた検索の結果に基づいて、第二塩基配列取得部3001により取得された塩基配列を遺伝子塩基配列にアラインメントする。例えば、部分塩基配列が符号2903で示される部分であるとして、部分塩基配列の類似塩基配列が符号2904で示される部分であることが、塩基配列検索装置3006により判明したとすると、符号2904で示される塩基配列の前後の塩基配列と、符号2902で示される塩基配列がどの程度似ているかを示すスコア値などを、ダイナミックプログラミングの手法などを用いて計算する。

[0139] (実施形態14:処理の流れ)

図31は、本実施形態に係る図30の塩基配列アラインメント装置の処理の流れを説明するフローチャートを例示する。ステップS3101において、第二塩基配列取得部3001により、塩基配列を取得する。ステップS3102において、部分塩基配列選択部3

002において、部分塩基配列を選択する。ステップS3103において、部分塩基配列入力部3003と第三ハミング距離入力部3004とにより、部分塩基配列とハミング距離を塩基配列検索装置3006へ入力する。ステップS3104により、塩基配列検索装置3006による検索の結果に基づいて塩基配列を遺伝子塩基配列にアラインメントする。ステップS3104は、ステップS3103で得られた検索の結果だけ繰り返して実行される。

[0140] （実施形態14：主な効果）

従来のアラインメントの手法では、BLASTなどが用いられていたが、BLASTなどを用いると、例えば連続する7merが一致する塩基配列の検索を行なって類似する塩基配列が遺伝子塩基配列のどこに出現するかを求めることになるので、アラインメントを正確に行なうことが困難な場合があった。本発明では、部分塩基配列の類似塩基配列を検索するので、より正確なアラインメントを行なうことができる。

産業上の利用可能性

[0141] 本発明に係る塩基配列検索装置及び塩基配列検索方法は、検索のために必要となる計算量を小さくすることができ、また、ハミング距離が所定の値以下となり、すなわち、似た塩基配列の存在を見落とすことも無いので、塩基配列などの設計に有用である。例えば、本発明に係る塩基配列検索装置及び塩基配列検索方法を、siRNAの塩基配列設計に適用した場合、特に、RNA干渉（RNAi）効果の高いsiRNAを設計可能とする種々の所定ガイドライン（具体的には、Ui-Teiらによるガイドライン Ui-Tei,K., Naito,Y., Takahashi,F., Haraguchi,T., Ohki-Hamazaki,H., Juni,A., Ueda,R. and Saigo,K., 'Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference', Nucleic Acids Research, 2004, Vol. 32, No.3, 936-948等）と組み合わせて適用した場合に、作業に要する時間の短縮及び適切な設計の点から、より有効である。

図面の簡単な説明

[0142] [図1]RNA干渉の過程の概略を示す図

[図2]マイクロアレイを用いた遺伝子解析や遺伝子診断などの過程を説明するための図

[図3]BLASTで見落とされる可能性のある塩基配列を説明するための図

[図4]本発明の実施形態1に係る塩基配列検索装置の機能ブロック図

[図5]ハミング距離が3の塩基配列の一例図

[図6]ハミング距離の定義を示す図

[図7]特定される2つの部分配列とその余の部分との一例図

[図8]ハミング距離の割り振りを説明するための図

[図9]割当部によるハミング距離の割り振りと選択部による選択を説明するための図

[図10]割当部によるハミング距離の割り振りと選択部による選択を説明するための図

[図11]本発明の実施形態1に係る塩基配列検索装置の処理の流れ図

[図12]本発明の実施形態2に係る塩基配列検索装置の機能ブロック図

[図13]本発明の実施形態3に係る塩基配列検索装置の機能ブロック図

[図14]本発明の実施形態4に係る塩基配列検索装置の機能ブロック図

[図15]本発明の実施形態5に係る塩基配列検索装置の機能ブロック図

[図16]本発明の実施形態6に係る塩基配列検索装置の機能ブロック図

[図17]ハミング距離を分割して割り当てるプログラムの一例図

[図18]置換塩基配列を生成するプログラムの一例図

[図19]本発明の実施形態10に係る塩基配列検索装置の機能ブロック図

[図20]リピート配列を格納する表の一例図

[図21]リピート配列情報を蓄積するための表の一例図

[図22]本発明の実施形態10に係る塩基配列検索装置の検索部の処理のフローチャート

[図23]本発明の実施形態11に係る塩基配列検索装置の機能ブロック図

[図24]入力塩基配列とハミング距離と類似塩基配列とを関連付けて蓄積するための表の構造図

[図25]本発明の実施形態11に係る塩基配列検索装置の判定部と類似塩基配列蓄積部との処理のフローチャート

[図26]本発明の実施形態12に係る塩基配列検索装置の機能ブロック図

[図27]本発明の実施形態13に係る無効果塩基配列生成装置の機能ブロック図

[図28]本発明の実施形態13に係る無効果塩基配列生成装置の処理のフローチャート

[図29]本発明の実施形態14における装置による処理の概要図

[図30]本発明の実施形態14に係る塩基配列アラインメント装置の機能ブロック図

[図31]本発明の実施形態14に係る塩基配列アラインメント装置の処理のフローチャート

符号の説明

[0143] 400 塩基配列検索装置

401 塩基配列入力部

402 ハミング距離入力部

403 特定部

404 割当部

405 選択部

406 置換塩基配列生成部

407 検索部

請求の範囲

[1] 遺伝子情報を表わす遺伝子塩基配列を格納したデータベースを検索するための索引であり、所定の長さである所定長の塩基配列が前記遺伝子塩基配列の中に出現する位置を検索するための索引、を用いて、入力される塩基配列と同じ長さで類似する塩基配列であり前記遺伝子塩基配列に出現する塩基配列である類似塩基配列を検索するための塩基配列検索装置であって、

前記所定長を超える長さの塩基配列を入力する塩基配列入力部と、

前記塩基配列入力部に入力された塩基配列である入力塩基配列に対して、適合しない塩基への置換の操作を行なう塩基の個数を示すハミング距離を入力するハミング距離入力部と、

前記入力塩基配列の部分配列であって、前記所定長の長さを持ち異なる2つの部分配列と、その余の部分と、を特定する特定部と、

前記特定部で特定された部分配列とその余の部分とに、前記ハミング距離入力部で入力されたハミング距離を分割して割り当てる割当部と、

前記特定部で特定された2つの部分配列のうち、前記割当部で割り当てられたハミング距離で示される個数の塩基を適合しない塩基へ置換する操作を前記部分配列に対して行なって生成される塩基配列である置換塩基配列の総数が大きくない方を選択する選択部と、

前記選択部により選択された部分配列に対して、前記割当部で割り当てられたハミング距離をもつ置換塩基配列を生成する置換塩基配列生成部と、

前記置換塩基配列生成部で生成された置換塩基配列を検索キーとして前記索引を用いて検索を行なう検索部と、

を有する塩基配列検索装置。

[2] 前記特定部は、

前記塩基配列入力部で入力された塩基配列の塩基数が前記所定長の2倍以下または2倍未満であれば、前記2つの部分配列のうちの一方の部分配列の端を前記入力塩基配列の一方の端と一致させ、前記2つの部分配列のうちの他方の部分配列の端を前記入力塩基配列の他方の端と一致させ、その余の部分が生じず特定されな

いことにする第一特定手段を有する請求項1に記載の塩基配列検索装置。

[3] 前記特定部は、

前記塩基配列入力部で入力された塩基配列の塩基数が前記所定長の2倍より大であれば、前記2つの部分配列が重ならないことにして前記2つの部分配列を特定する第二特定手段を有する請求項1または2に記載の塩基配列検索装置。

[4] 前記検索部での検索結果に基づいて、前記置換塩基配列を含んで遺伝子塩基配列に現れる塩基配列である類似候補塩基配列を取得する類似候補塩基配列取得部と、

前記類似候補塩基配列取得部で取得された類似候補塩基配列と前記入力塩基配列とのハミング距離が前記ハミング距離入力部に入力されたハミング距離と同じ、又はそれ未満であるかどうかを判定する判定部と、

を有する請求項1から3のいずれか一に記載の塩基配列検索装置。

[5] 適合しない塩基の組を指定する不適合塩基組入力部を有し、不適合塩基組入力部に入力された塩基の組に基づいて、検索部で検索が行なわれ、また、ハミング距離が求められる請求項4に記載の塩基配列検索装置。

[6] 前記塩基配列入力部に入力された塩基配列と類似塩基配列との対応する塩基の適合の分布を表わす分布情報を入力する適合分布入力部を有し、

前記判定部は、前記適合分布入力部で入力された分布情報が満たされているかどうかを判定する分布判定手段を有する請求項4または5のいずれか一に記載の塩基配列検索装置。

[7] 前記適合分布入力部で入力される分布情報は、塩基配列と類似塩基配列との対応する塩基が連続して適合する長さの下限である請求項6に記載の塩基配列検索装置。

[8] 前記塩基配列入力部に入力される塩基配列の長さは15から60であり、前記所定長は、11から14である請求項1から7のいずれか一に記載の塩基配列検索装置。

[9] 遺伝子情報を表わす遺伝子塩基配列を格納したデータベースを検索するための索引であって、所定の長さである所定長の塩基配列が前記遺伝子塩基配列の中に出現する位置を検索するための索引、を用いて、入力される塩基配列と同じ長さで

類似する塩基配列であり前記遺伝子塩基配列に出現する塩基配列である類似塩基配列を検索するための塩基配列検索方法であって、

前記所定長を超えるの長さの塩基配列を入力する塩基配列入力ステップと、

前記塩基配列入力部に入力された塩基配列である入力塩基配列に対して、適合しない塩基への置換の操作を行なう塩基の個数を示すハミング距離を入力するハミング距離入力ステップと、

前記入力塩基配列の部分配列であって、前記所定長の長さを持ち異なる2つの部分配列と、その余の部分と、を特定する特定ステップと、

前記特定ステップで特定された2つの部分配列とその余の部分とに、前記ハミング距離入力ステップにて入力されたハミング距離を分割して割り当てる割当ステップと、

前記特定ステップで特定された2つの部分配列のうち、前記割当部で割り当てられたハミング距離で示される個数の塩基を適合しない塩基へ置換する操作を前記部分配列に対して行なって生成される塩基配列である置換塩基配列の総数が大きくない方を選択する選択ステップと、

前記選択ステップにより選択された部分配列に対して、前記割当ステップにて割り当てられたハミング距離をもつ置換塩基配列を生成する置換塩基配列生成ステップと、

前記置換塩基配列生成ステップで生成された部分配列を検索キーとして前記索引を用いて検索を行なう検索ステップと、

を含む塩基配列検索方法。

[10] アルファベットが一次元に配列した文字列を格納したデータベースを検索するための索引であり、所定の長さである所定長の文字列が前記データベースに格納された文字列の中に出現する位置を検索するための索引、を用いて、入力される文字列と同じ長さで類似する文字列であり前記前記データベースに格納された文字列に出現する文字列である類似文字列を検索するための文字列検索装置であって、

前記所定長を超える長さの文字列を入力する文字列入力部と、

前記文字列入力部に入力された文字列である入力文字列に対して、適合しないアルファベットへの置換の操作を行なうアルファベットの個数を示すハミング距離を入力

するハミング距離入力部と、

前記入力文字列の部分文字列であって、前記所定長の長さを持ち異なる2つの部分文字列と、その余の部分と、を特定する特定部と、

前記特定部で特定された部分文字列とその余の部分とに、前記ハミング距離入力部で入力されたハミング距離を分割して割り当てる割当部と、

前記特定部で特定された2つの部分文字列のうち、前記割当部で割り当てられたハミング距離で示される個数のアルファベットを適合しないアルファベットへ置換する操作を前記部分文字列に対して行なって生成される文字列である置換文字列の総数が大きくない方を選択する選択部と、

前記選択部により選択された部分文字列に対して、前記割当部で割り当てられたハミング距離をもつ置換文字列を生成する置換文字列生成部と、

前記置換文字列生成部で生成された置換文字列を検索キーとして前記索引を用いて検索を行なう検索部と、

を有する文字列検索装置。

[11] 前記文字列は、ペプチド配列である請求項10に記載の文字列検索装置。

[12] 遺伝子塩基配列中に繰り返して出現する前記所定長の塩基配列を蓄積するリピート配列蓄積部と、

前記リピート配列蓄積部に蓄積された塩基配列に、その塩基配列の前記遺伝子塩基配列中における出現位置を関連付けた情報であるリピート配列情報を蓄積するリピート配列情報蓄積部と、

を有し、

前記検索部は、

前記置換塩基配列が前記リピート配列蓄積部に蓄積されているかどうかを判定するリピート配列判定手段と、

前記リピート配列判定手段にて前記置換塩基配列が前記リピート配列蓄積部に蓄積されていると判定された場合には、前記リピート配列情報蓄積部に蓄積されたリピート配列情報に基づいて検索を行なうリピート配列検索手段と、

を有する請求項1から8のいずれか一に記載の塩基配列検索装置。

[13] 前記判定部にて、前記入力塩基配列と、前記類似候補塩基配列取得部により取得された類似候補塩基配列と、のハミング距離が前記ハミング距離入力部に入力されたハミング距離以下であると判定された場合に、前記入力塩基配列と、前記入力塩基配列と前記類似塩基配列とのハミング距離と、前記類似候補塩基配列と、を関連付けて蓄積する類似塩基配列蓄積部を有する請求項4から7のいずれか一に記載の塩基配列検索装置。

[14] 前記判定部にて、前記類似候補塩基配列取得部により取得された類似候補塩基配列と前記入力塩基配列とのハミング距離が前記ハミング距離入力部に入力されたハミング距離以下であると判定された場合に、前記塩基配列入力部により入力された塩基配列と前記類似候補塩基配列取得部で取得された類似候補塩基配列の会合率を計算する会合率計算部

を有する請求項4から7のいずれか一に記載の塩基配列検索装置。

[15] 前記所定長を超える長さの塩基配列を取得する塩基配列取得部と、

前記塩基配列取得部で取得された塩基配列の塩基のうち、所定の個数の塩基を置換して得られる塩基配列である無効果候補置換塩基配列を生成する無効果候補置換塩基配列生成部と、

前記無効果候補置換塩基配列生成部で生成された無効果候補置換塩基配列を請求項14に記載の塩基配列検索装置に入力する無効果候補置換塩基配列入力部と、

所定のハミング距離を前記無効果候補置換塩基配列入力部が無効果候補置換塩基配列を入力した塩基配列検索装置に入力する第二ハミング距離入力部と、

前記無効果候補置換塩基配列生成部で生成された無効果候補置換塩基配列の中から、前記無効果候補置換塩基配列入力部による入力と前記第二ハミング距離入力部による入力とによって前記塩基配列検索装置より得られた会合率の低い塩基配列を選択する選択部と、

を備える無効果塩基配列生成装置。

[16] 前記所定長を超える長さの塩基配列を取得する第二塩基配列取得部と、

前記第二塩基配列取得部で取得された塩基配列の一部分である部分塩基配列を

選択する部分塩基配列選択部と、

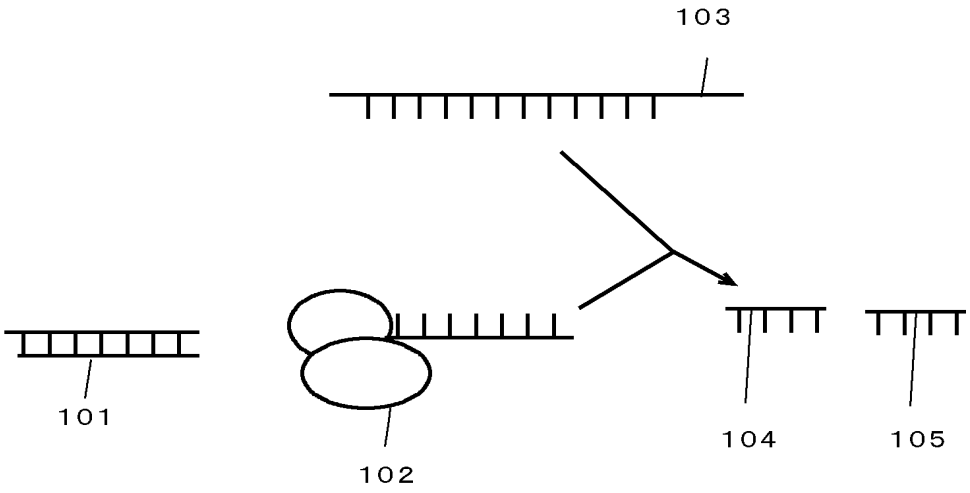
前記部分塩基配列選択部で選択された部分塩基配列を請求項4から8のいずれか一に記載の塩基配列検索装置に入力する部分塩基配列入力部と、

所定のハミング距離を前記部分塩基配列入力部が部分塩基配列を入力した塩基配列検索装置に入力する第三ハミング距離入力部と、

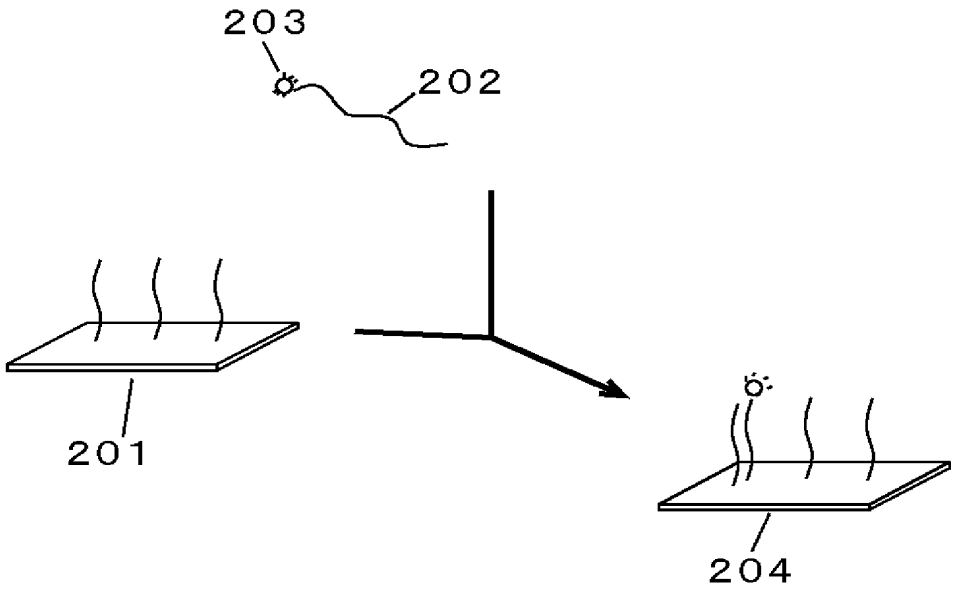
前記部分塩基配列入力部による入力と前記第三ハミング距離入力部によるの入力とによって前記塩基配列検索装置より得られた検索の結果に基づいて、前記第二塩基配列取得部により取得された塩基配列を前記遺伝子塩基配列にアラインメントするアラインメント部と、

を有する塩基配列アラインメント装置。

[図1]



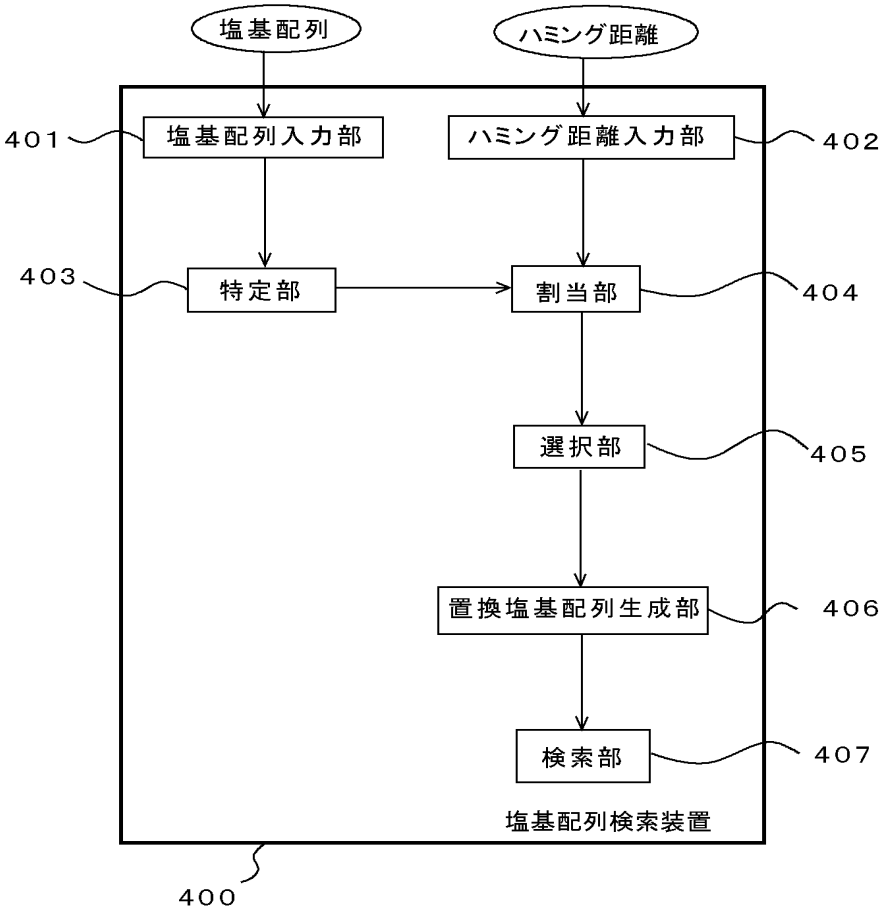
[図2]



[図3]

○○○○○×○○○○○○×○○○○○○
○○○○○○×○○○○○×○○○○○○
○○○○○○×○○○○○○×○○○○○

[図4]



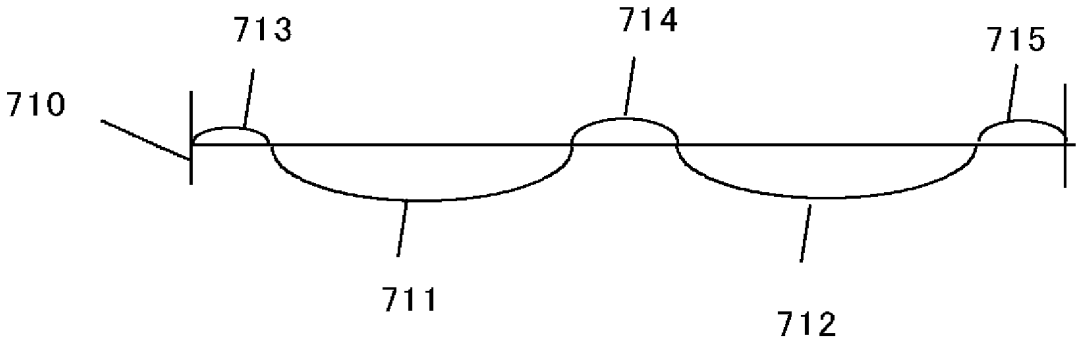
[図5]

[図6]

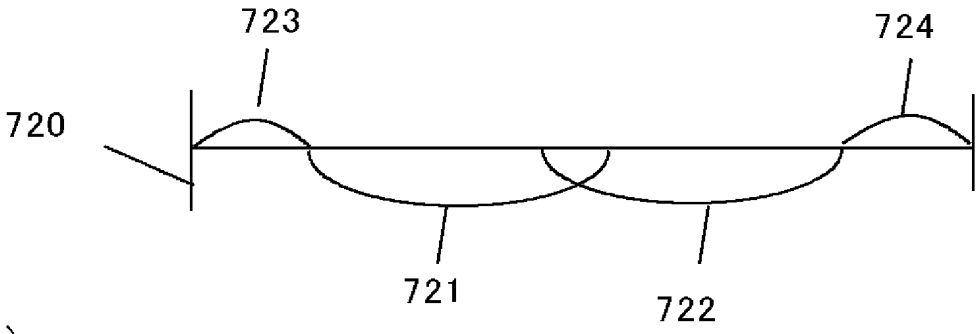
$$d_H(S,T)= \left| \{ i \mid S_i \neq T_i,\ i=1,2,\cdots,n \} \right|$$

[図7]

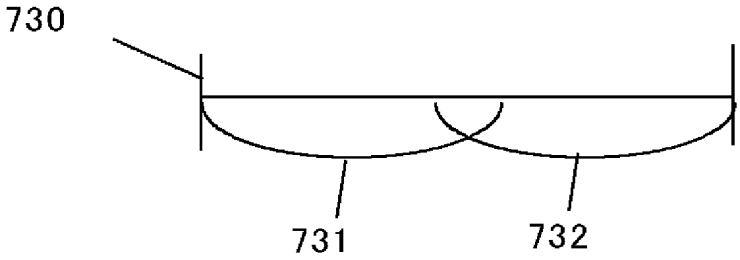
（1）


713
714
715
710
711
712

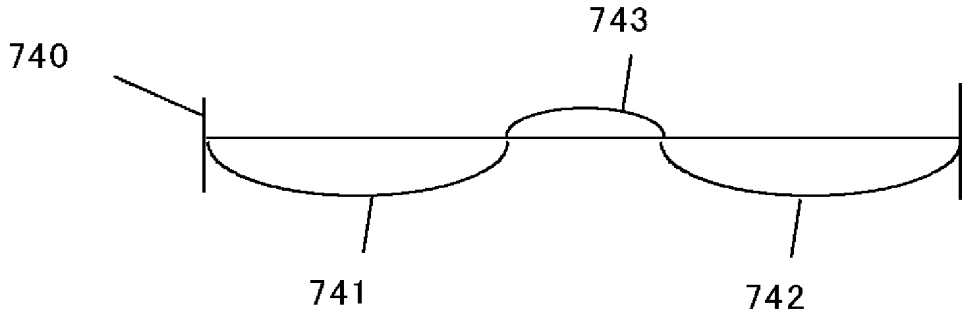
（2）


723
724
720
721
722

（3）

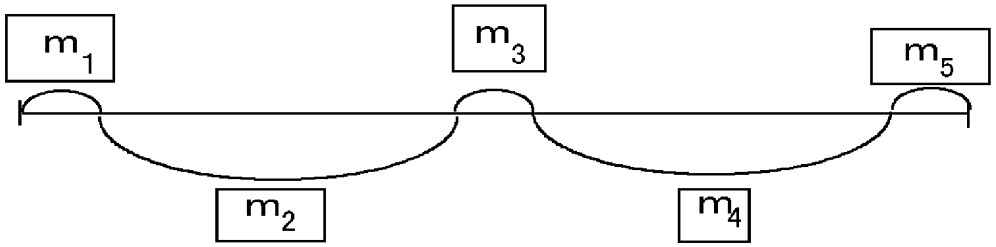

730
731
732

（4）


743
740
741
742

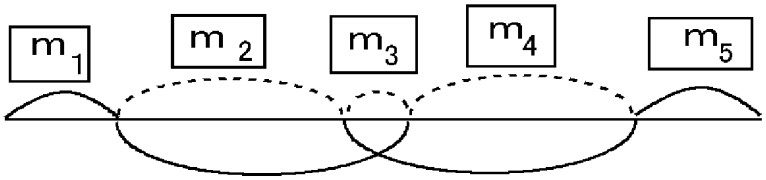
[図8]

（１）



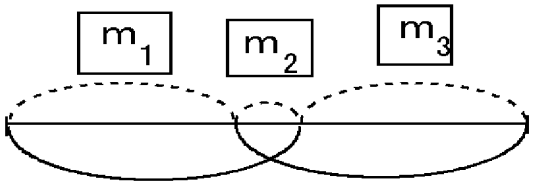
$m_1 + m_2 + m_3 + m_4 + m_5$ ＝入力ハミング距離

（２）



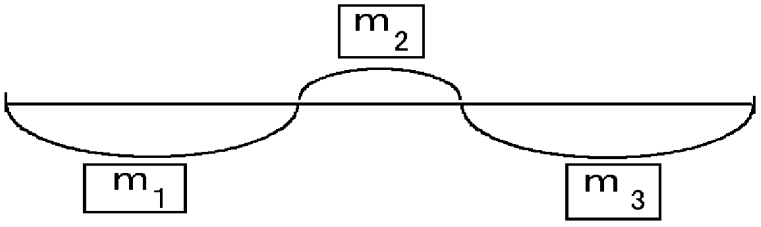
$m_1 + m_2 + m_3 + m_4 + m_5$ ＝入力ハミング距離

（３）



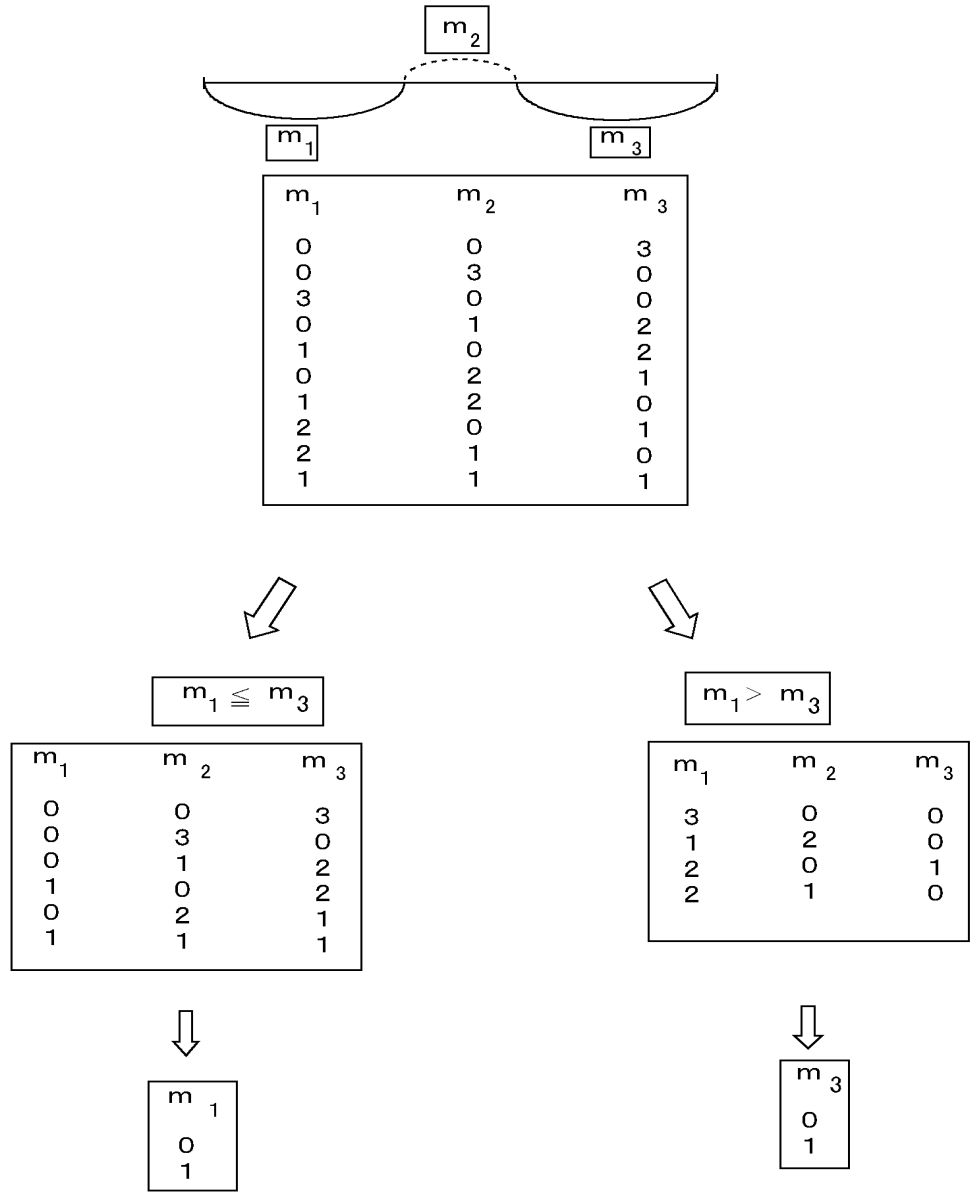
$m_1 + m_2 + m_3$ ＝ 入力ハミング距離

（４）

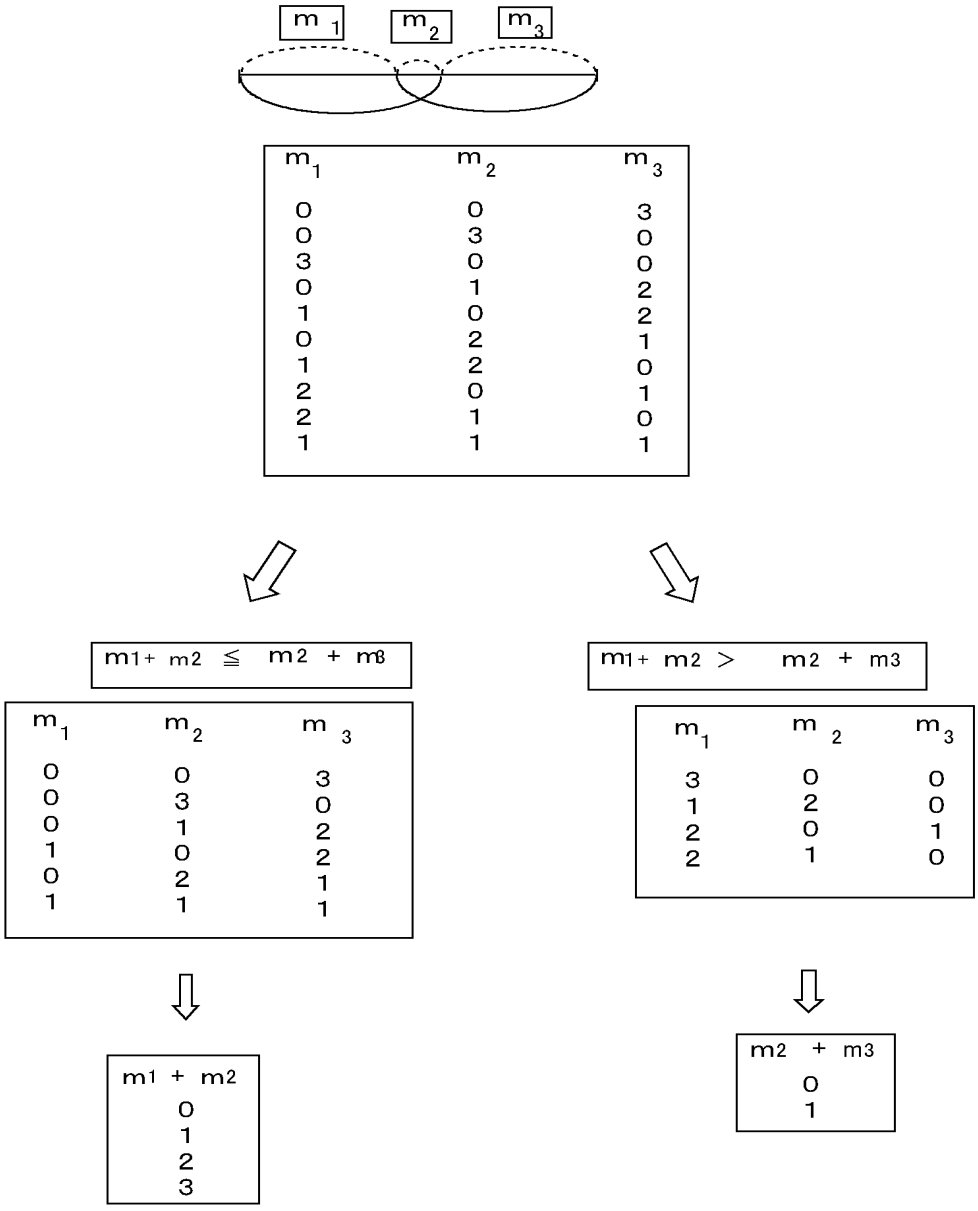


$m_1 + m_2 + m_3$ ＝ 入力ハミング距離

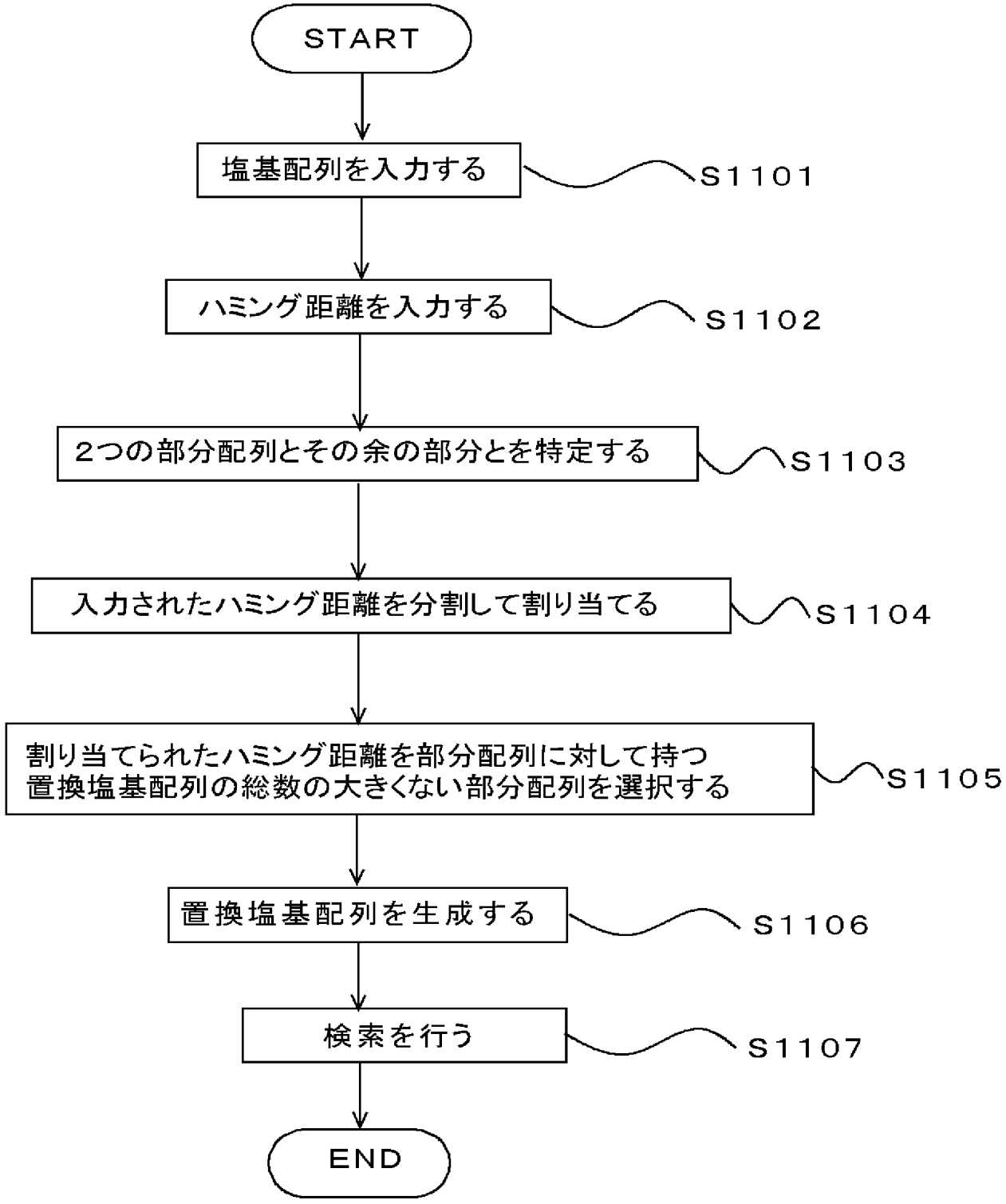
[図9]


m2
m1
m3
m1 m2 m3
0 0 3
0 3 0
3 0 0
0 1 2
1 0 2
0 2 1
1 2 0
2 0 1
2 1 0
1 1 1
m1 ≦ m3
m1 m2 m3
0 0 3
0 3 0
0 1 2
1 0 2
0 2 1
1 1 1
m1
0
1
m1 > m3
m1 m2 m3
3 0 0
1 2 0
2 0 1
2 1 0
m3
0
1

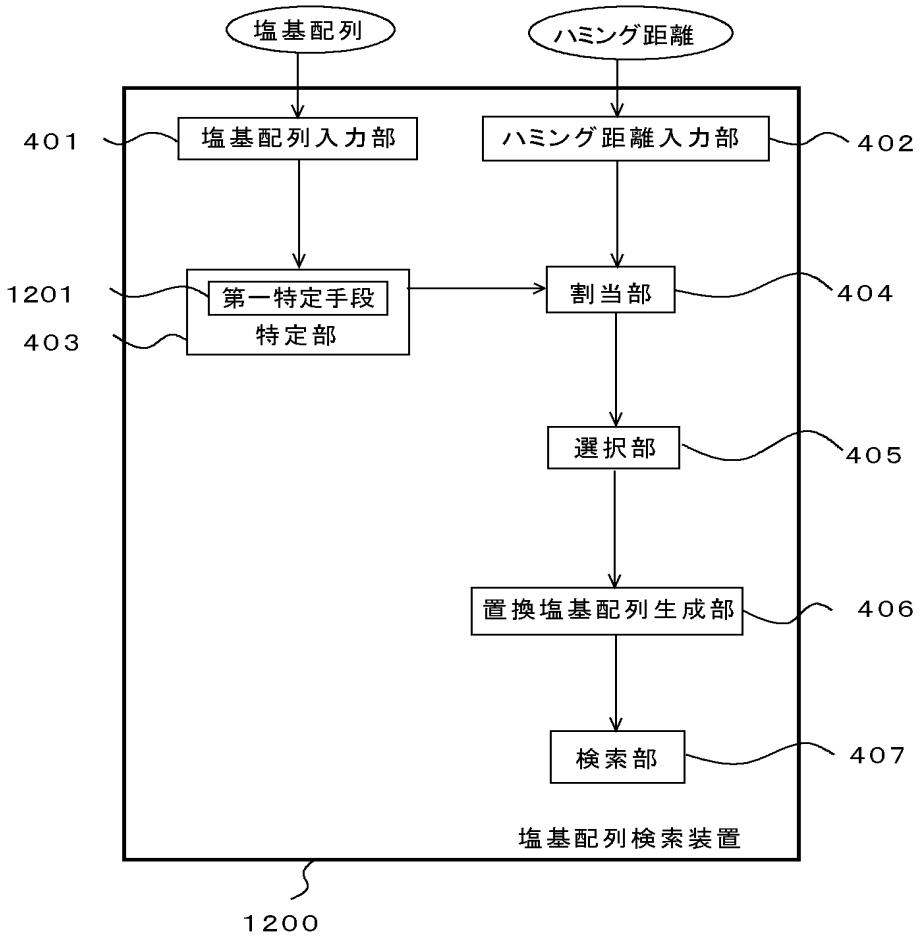
[図10]


m1
m2
m3
m1 m2 m3
0 0 3
0 3 0
3 0 0
0 1 2
1 0 2
0 2 1
1 2 0
2 0 1
2 2 0
2 1 0
1 1 1
m1+ m2 ≦ m2 + m3
m1 m2 m3
0 0 3
0 3 0
0 1 2
1 0 2
0 2 1
1 1 1
m1 + m2
0
1
2
3
m1+ m2 > m2 + m3
m1 m2 m3
3 0 0
1 2 0
2 0 1
2 1 0
m2 + m3
0
1

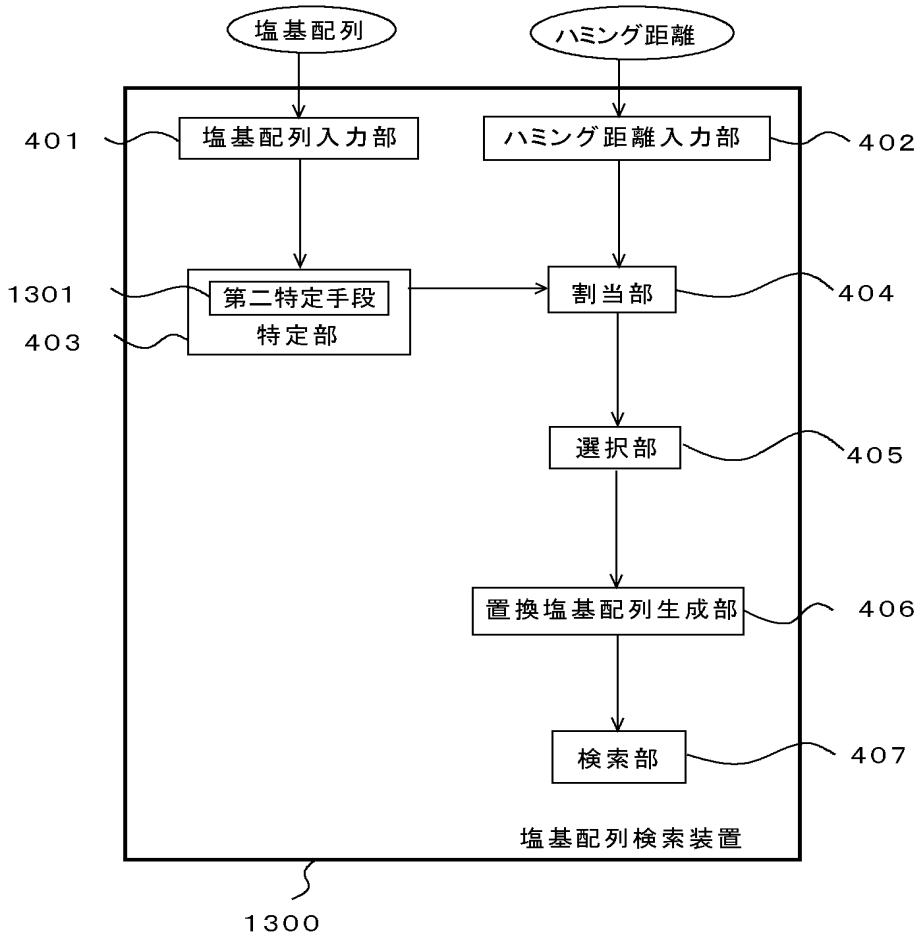
[図11]


START
塩基配列を入力する
S1101
ハミング距離を入力する
S1102
2つの部分配列とその余の部分とを特定する
S1103
入力されたハミング距離を分割して割り当てる
S1104
割り当てられたハミング距離を部分配列に対して持つ
置換塩基配列の総数の大きくない部分配列を選択する
S1105
置換塩基配列を生成する
S1106
検索を行う
S1107
END

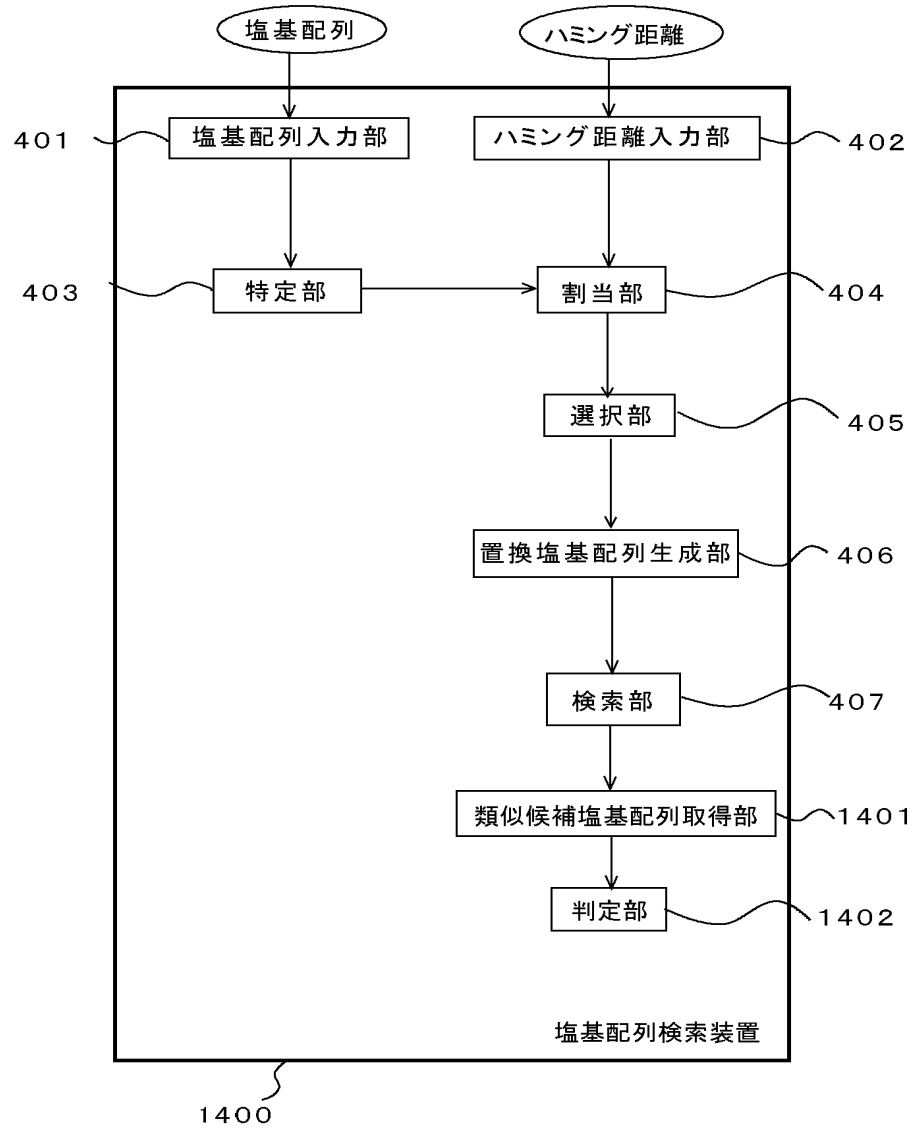
[図12]


塩基配列
ハミング距離
401
塩基配列入力部
ハミング距離入力部
402
1201
第一特定手段
403
特定部
割当部
404
選択部
405
置換塩基配列生成部
406
検索部
407
塩基配列検索装置
1200

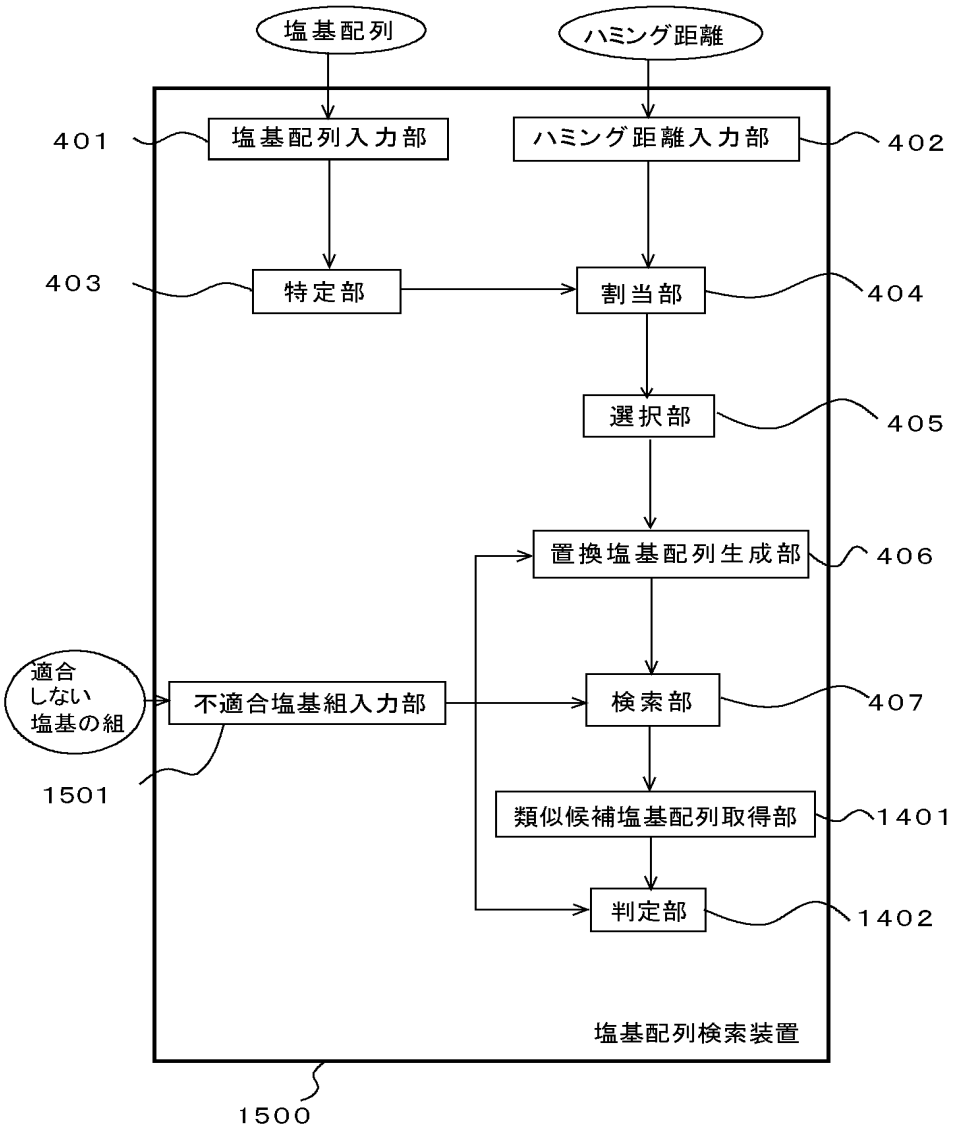
[図13]


塩基配列
ハミング距離
401
塩基配列入力部
ハミング距離入力部
402
1301
第二特定手段
403
特定部
割当部
404
選択部
405
置換塩基配列生成部
406
検索部
407
塩基配列検索装置
1300

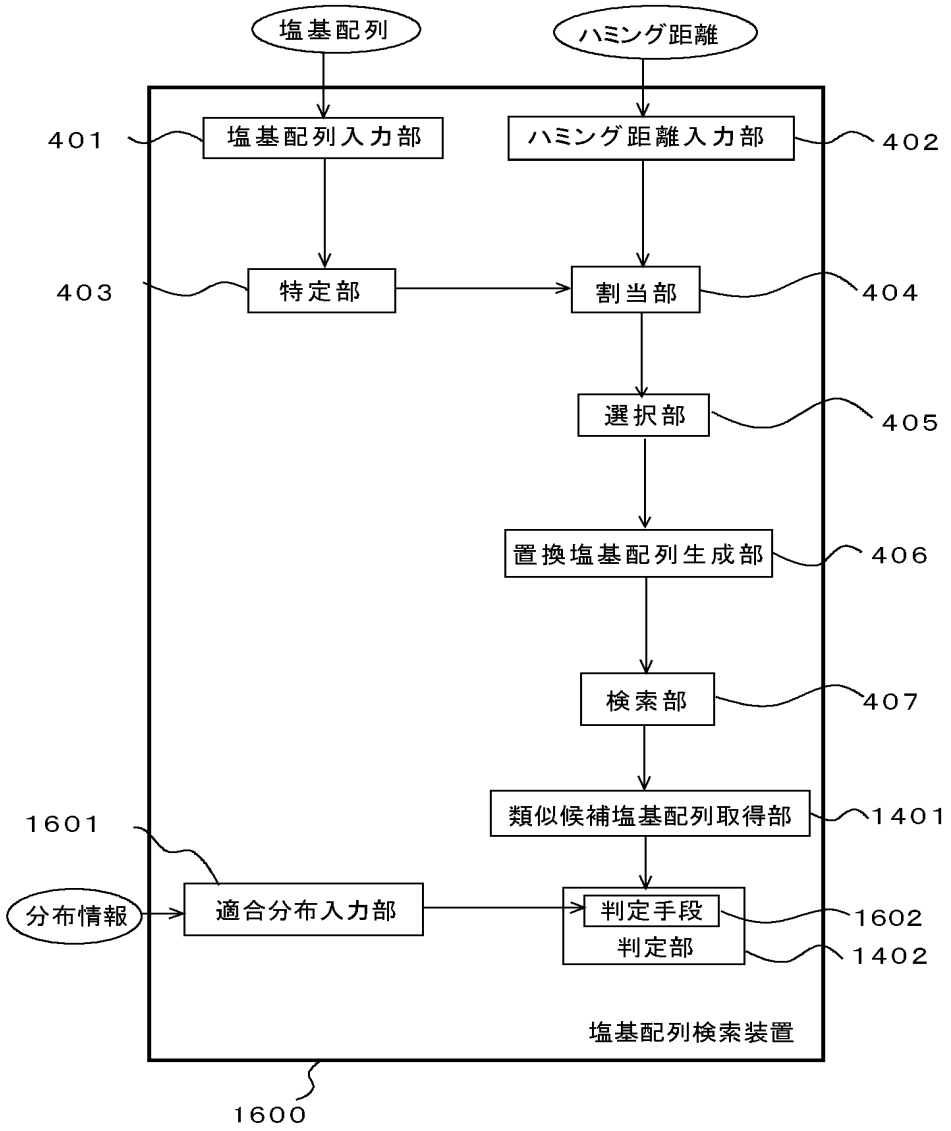
[図14]


塩基配列
ハミング距離
401
塩基配列入力部
ハミング距離入力部
402
403
特定部
割当部
404
選択部
405
置換塩基配列生成部
406
検索部
407
類似候補塩基配列取得部
1401
判定部
1402
塩基配列検索装置
1400

[図15]


塩基配列
ハミング距離
401
塩基配列入力部
ハミング距離入力部
402
403
特定部
割当部
404
選択部
405
置換塩基配列生成部
406
適合
しない
塩基の組
不適合塩基組入力部
検索部
407
1501
類似候補塩基配列取得部
1401
判定部
1402
塩基配列検索装置
1500

[図16]


塩基配列
ハミング距離
401
塩基配列入力部
ハミング距離入力部
402
403
特定部
割当部
404
選択部
405
置換塩基配列生成部
406
検索部
407
類似候補塩基配列取得部
1401
1601
分布情報
適合分布入力部
判定手段
判定部
1602
1402
塩基配列検索装置
1600

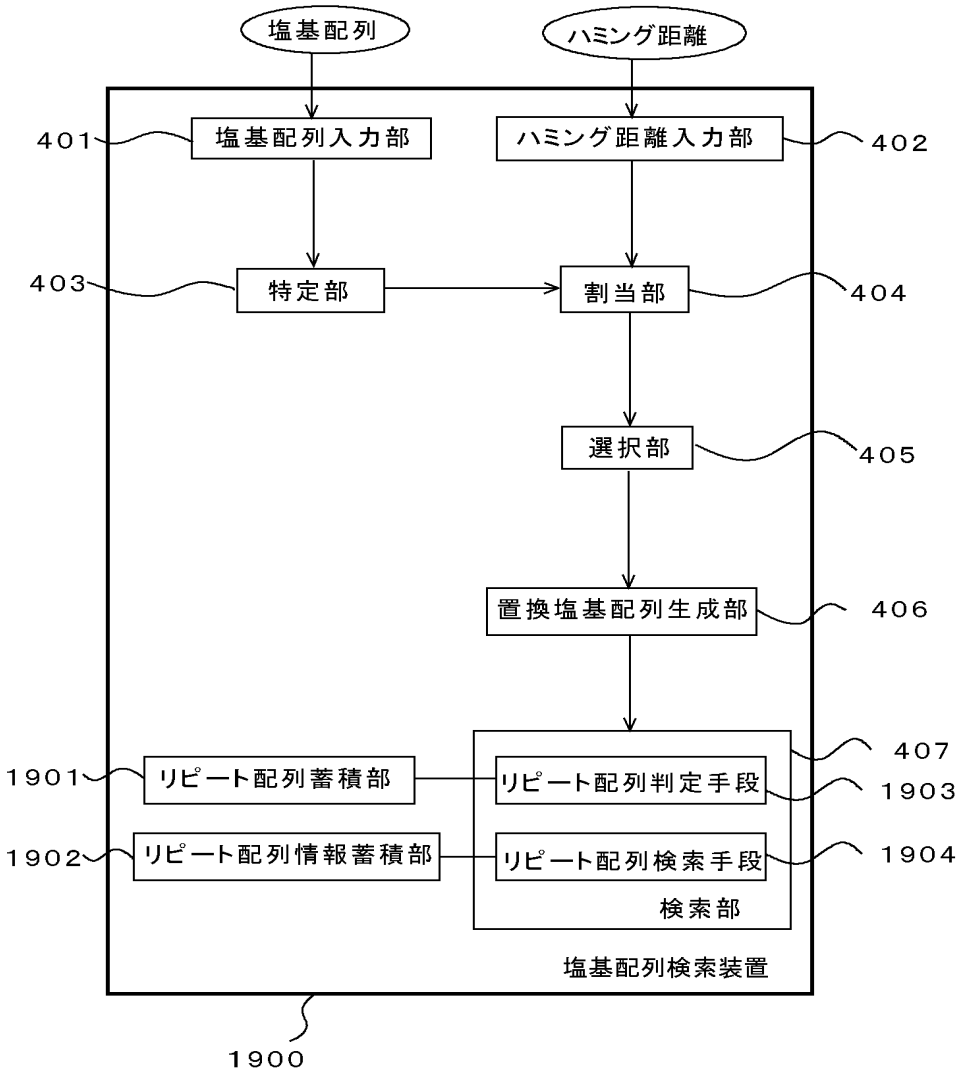
[図17]

```
distributeHammingDistance(int P, int H, int nSize, int* vec)
{
    int h;

    if (P==1) {
        vec[1] = h;

        /*
         * 全ての部分へのハミング距離の割り当ての一つが
         * 完成したので vec に格納されたハミング距離を出力する
         */
        for (int i = 1; i <= nSize; i = i + 1) {
            printf("Part %d th: %d", i, vec[i]);
            /* セパレータ又はターミネータを出力する */
            if (i != nSize) {
                /* セパレータとしてカンマを出力する */
                printf(", ");
            }
            else {
                /* ターミネータとして改行を出力する */
                printf("¥n");
            }
        }

    }
    else {
        for (h = 0; h <= H; h = h + 1) {
            vec[P] = h;
            distributeHammingDistance(P - 1, H - h, nSize, vec);
        }
    }
}
```

[図18]

```
for (I1 = 0; I1 < L; I1 = I1 + 1) {
    for (I2 = I1 + 1; I2 < L; I2 = I2 + 1) {
        foreach a1 in {A, C, G, T} {
            if (S[I1] != a1 ) {
                foreach a2 in {A, C, G, T} {
                    if (S[I2] != a2) {
                        S のI1番目の塩基をa1に置換し、
                        SのI2番目の塩基をa2に置換して
                        得られる置換塩基配列を生成 ;
                    }
                }
            }
        }
    }
}
```

[図19]



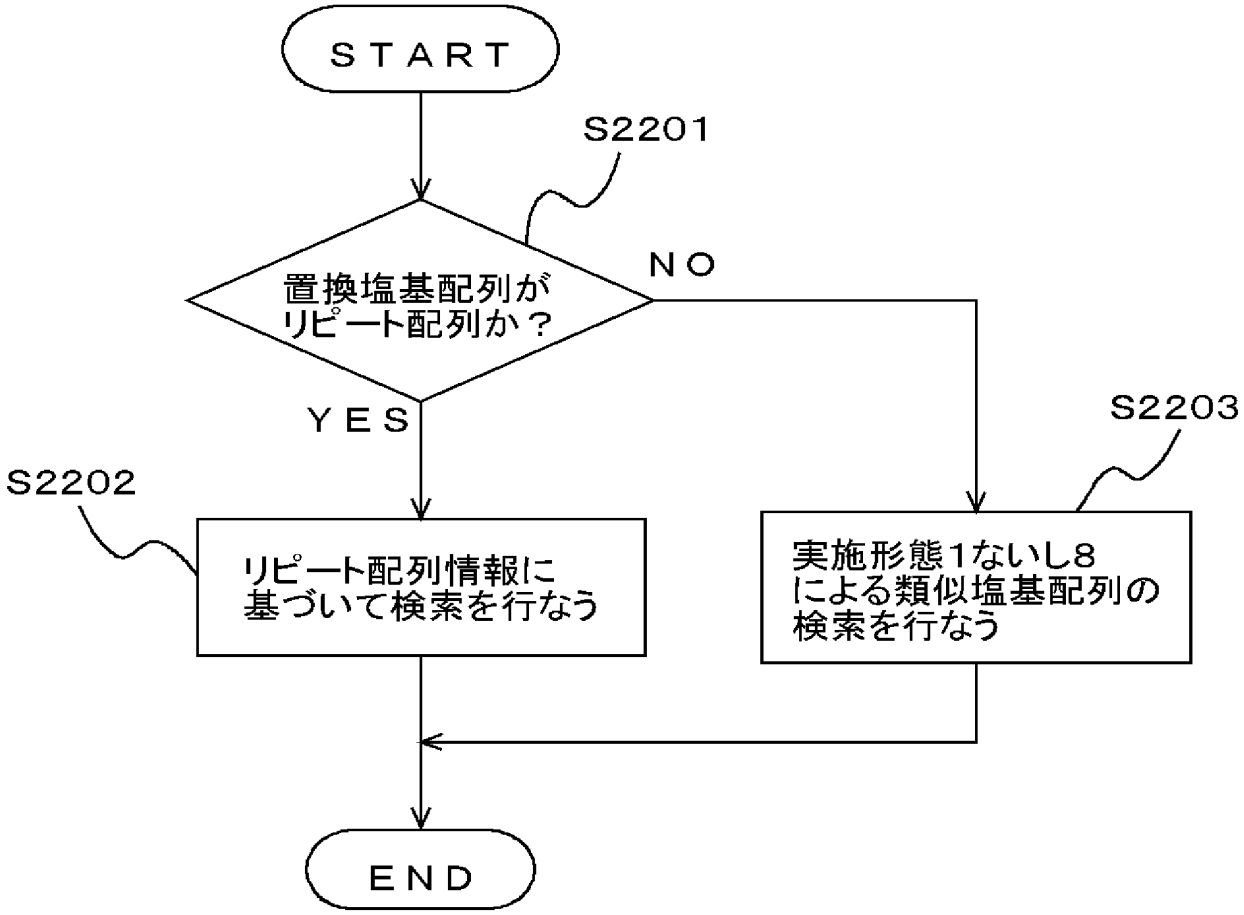
[図20]

リピート配列識別子	リピート配列
1	ACGGGUC・・・
2	GGGGGAAAA・・・
⋮	⋮

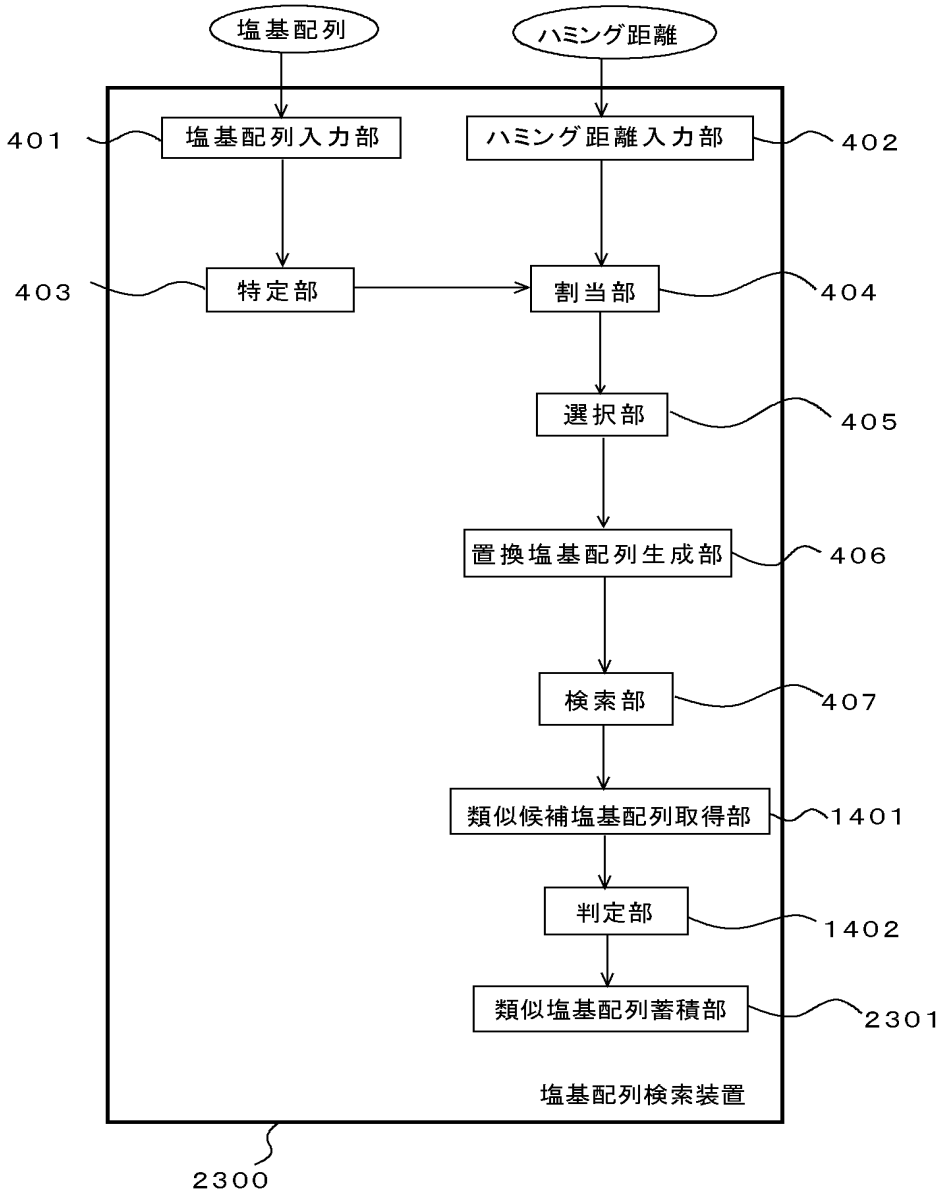
[図21]

リピート配列識別子	出現位置
1	58210
1	37703
1	27503
1	30516
⋮	⋮
2	167
2	27367
2	112109
⋮	⋮

[図22]


START
S2201
置換塩基配列が
リピート配列か？
NO
YES
S2202
リピート配列情報に
基づいて検索を行なう
S2203
実施形態1ないし8
による類似塩基配列の
検索を行なう
END

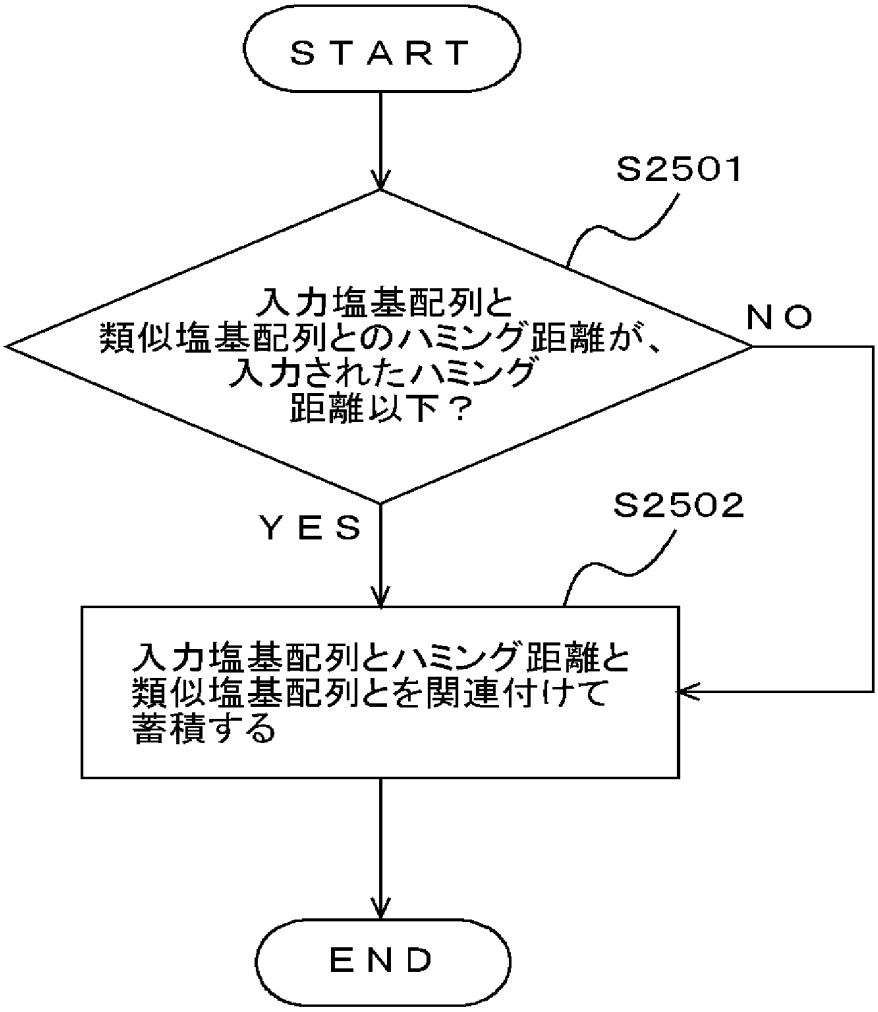
[図23]



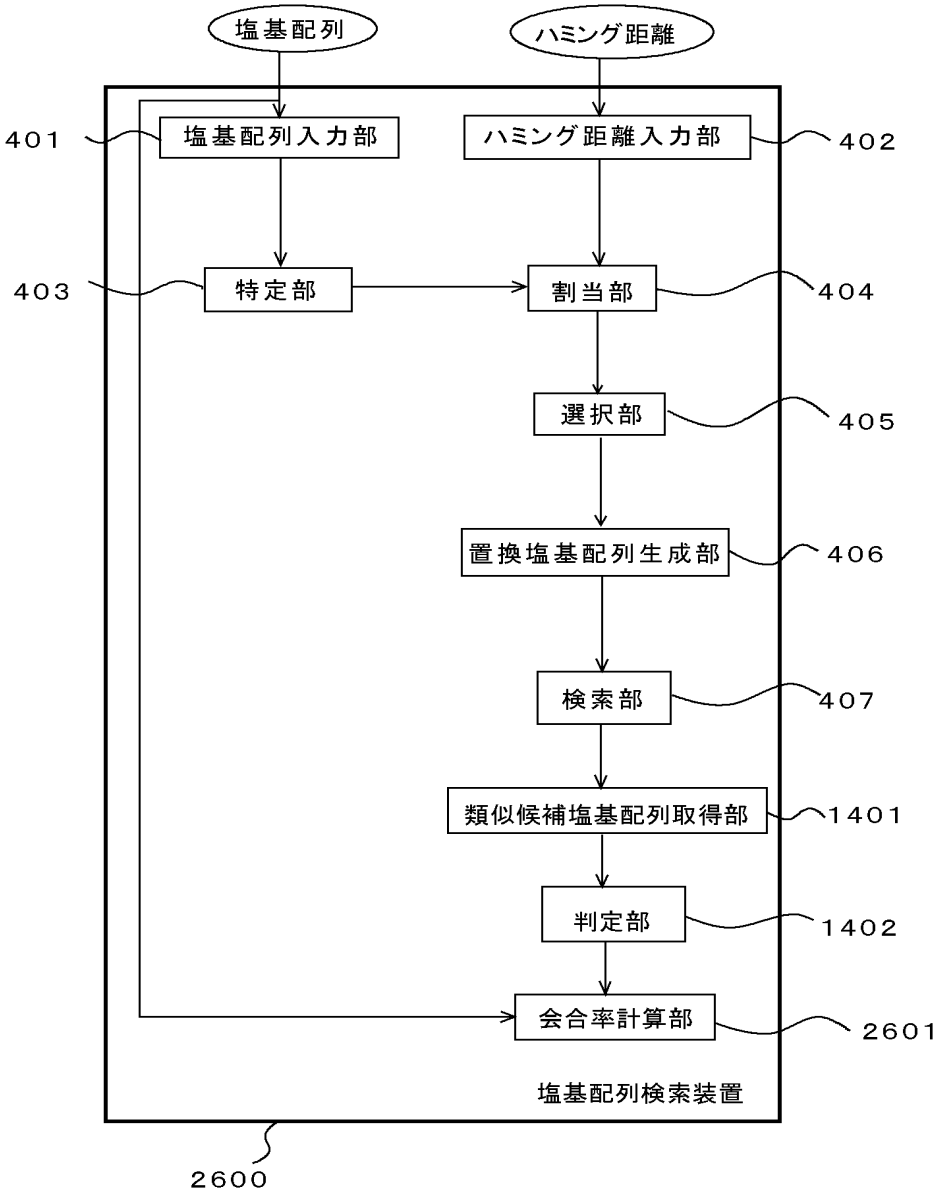
[図24]

入力塩基配列	ハミング距離	類似塩基配列
⋮	⋮	⋮

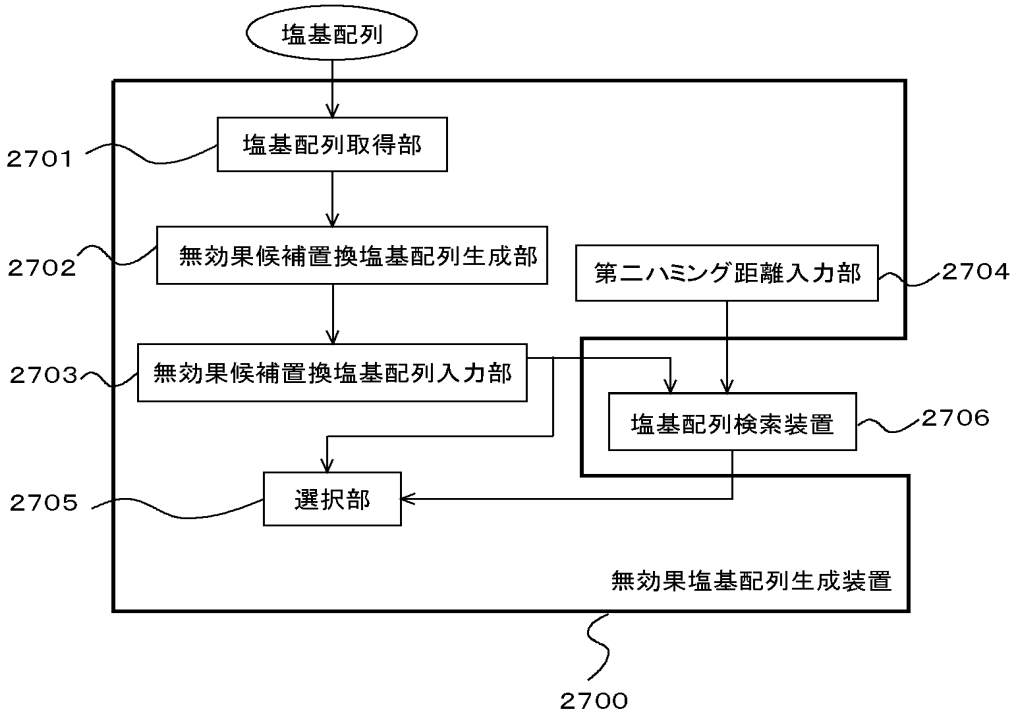
[図25]


START
S2501
入力塩基配列と
類似塩基配列とのハミング距離が、
入力されたハミング
距離以下？
NO
YES
S2502
入力塩基配列とハミング距離と
類似塩基配列とを関連付けて
蓄積する
END

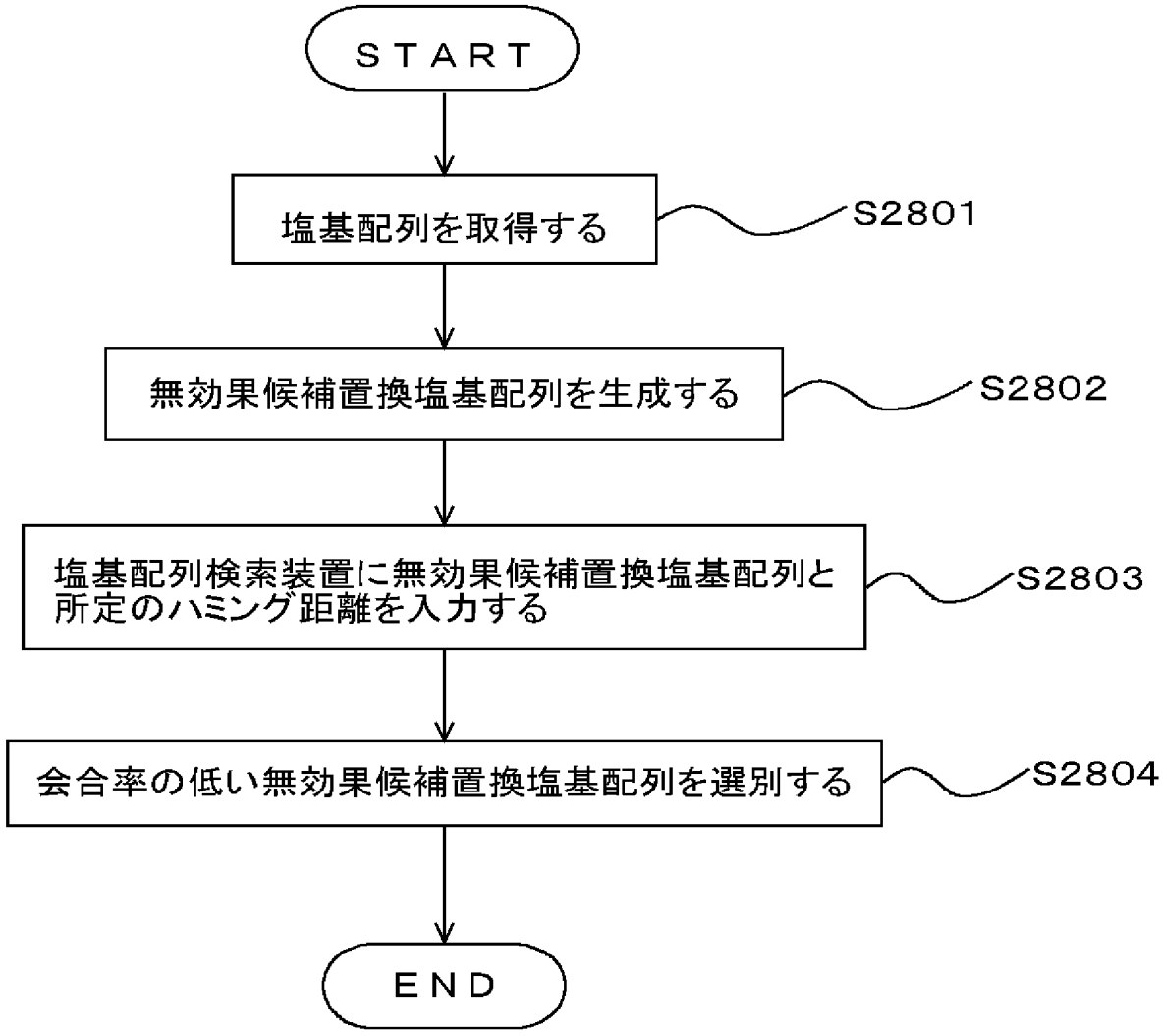
[図26]


塩基配列
ハミング距離
401
塩基配列入力部
ハミング距離入力部
402
403
特定部
割当部
404
選択部
405
置換塩基配列生成部
406
検索部
407
類似候補塩基配列取得部
1401
判定部
1402
会合率計算部
2601
塩基配列検索装置
2600

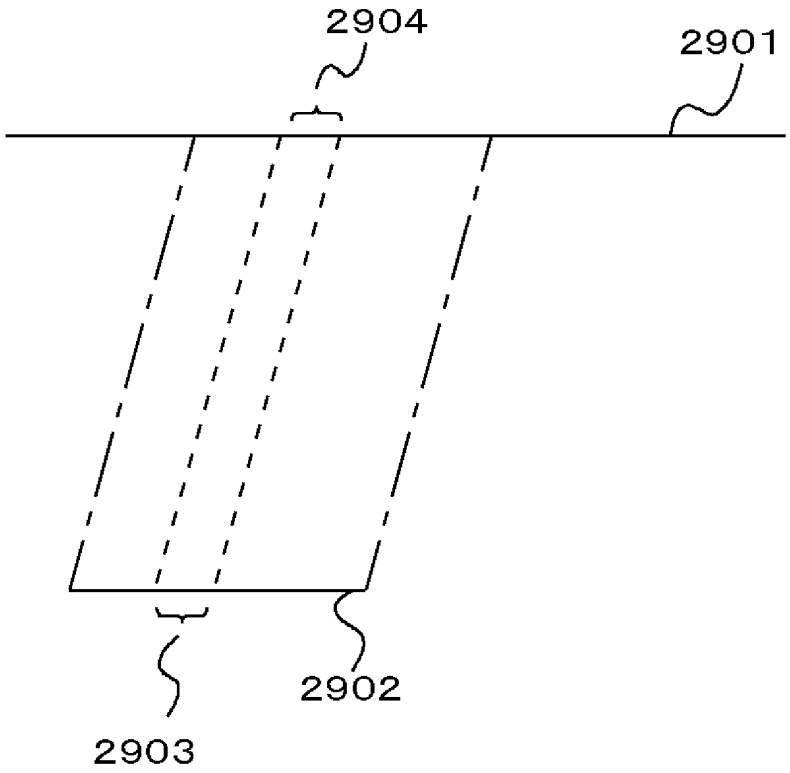
[図27]


塩基配列
2701
塩基配列取得部
2702
無効果候補置換塩基配列生成部
第二ハミング距離入力部
2704
2703
無効果候補置換塩基配列入力部
塩基配列検索装置
2706
2705
選択部
無効果塩基配列生成装置
2700

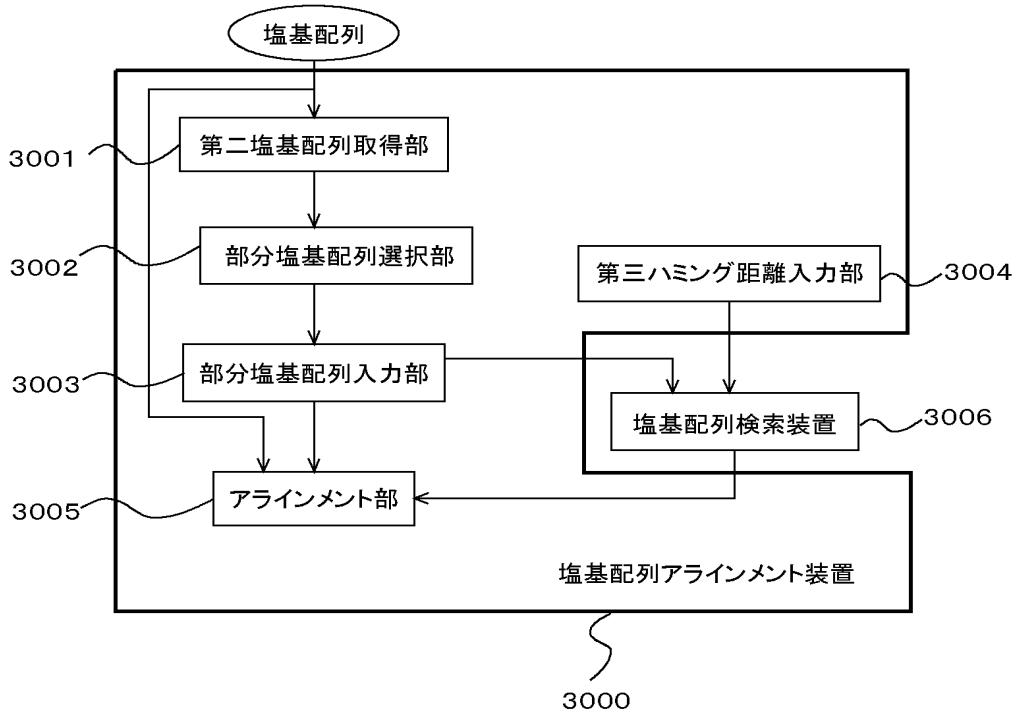
[図28]


START
塩基配列を取得する
S2801
無効果候補置換塩基配列を生成する
S2802
塩基配列検索装置に無効果候補置換塩基配列と
所定のハミング距離を入力する
S2803
会合率の低い無効果候補置換塩基配列を選別する
S2804
END

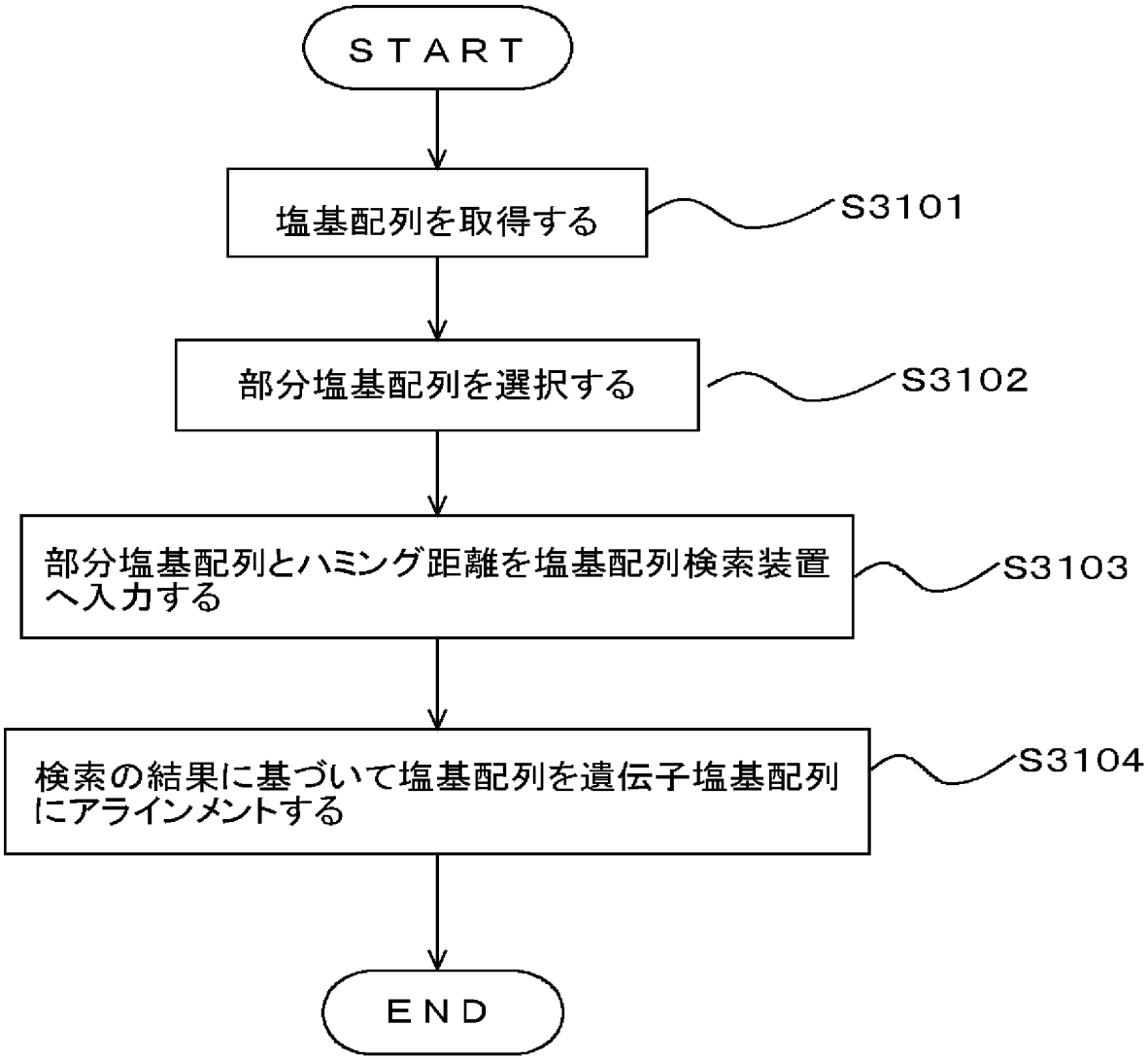
[図29]


2904
2901
2902
2903

[図30]


塩基配列
第二塩基配列取得部
3001
部分塩基配列選択部
3002
第三ハミング距離入力部
3004
部分塩基配列入力部
3003
塩基配列検索装置
3006
アラインメント部
3005
塩基配列アラインメント装置
3000

[図31]


START
塩基配列を取得する
S3101
部分塩基配列を選択する
S3102
部分塩基配列とハミング距離を塩基配列検索装置へ入力する
S3103
検索の結果に基づいて塩基配列を遺伝子塩基配列にアラインメントする
S3104
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2005/006397

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F19/00, 17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F19/00, 17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2005
Kokai Jitsuyo Shinan Koho 1971-2005 Toroku Jitsuyo Shinan Koho 1994-2005

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus(JOIS), PubMed

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	UI-TEI K et al., Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference, Nucleic Acids Research, 09 February, 2004 (09.02.04), Vol.32, No.3, pages 936 to 948	1-16
A	LIM et al., Finding Similar Regions In Many Strings, Proc.Annu.ACM Symp. Theory Comput, 1999, Vol.31, pages 473 to 482	1-16
A	Navarro G, A Guided Tour to Approximate String Matching, ACM Computing Surveys, 2001, Vol.33, No.1, pages 31 to 88	1-16

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 May, 2005 (17.05.05)	07 June, 2005 (07.06.05)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００５／００６３９７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.[7]　G06F19/00, 17/30

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.[7]　G06F19/00, 17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００５年
日本国実用新案登録公報	１９９６－２００５年
日本国登録実用新案公報	１９９４－２００５年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus(JOIS), PubMed

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	UI-TEI K, et al., Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference, Nucleic Acids Research, 2004.02.09, Vol. 32, No.3, pp. 936-948	1-16
A	LI M, et al, Finding Similar Regions In Many Strings, Proc Annu ACM Symp Theory Comput, 1999, Vol. 31st, pp.473-482	1-16
A	Navarro G, A Guided Tour to Approximate String Matching, ACM Computing Surveys, 2001, Vol.33, No.1, pp.31-88	1-16

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日 １７．０５．２００５	国際調査報告の発送日　07. 6. 2005
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　5L　9287 岩間　直純 電話番号　０３－３５８１－１１０１　内線　３５６２

様式ＰＣＴ／ＩＳＡ／２１０（第２ページ）（２００４年１月）